Comparing Meaning in Language and Cognition:

P-Hyponymy, Concept Combination, Asymmetric Similarity

Candidate number: 123 893

University of Oxford

Desislava Bankova

A thesis submitted for the degree of

MSc in Mathematics and Foundations of Computer Science

Trinity 2015

This thesis is dedicated to
my parents Albena and Yordan.

Abstract

In this dissertation we work in the framework of compositional distributional models of meaning to examine a number of asymmetric linguistic phenomena that manifest themselves in language and cognition. These include overextension with respect to concept combination, asymmetry of similarity judgment and hyponymy and typicality. In particular, we make use of the formalism of density matrices, which were recently introduced as an alternative to the vector-based model of word meaning. We first consider the former two of the above-mentioned phenomena using only tools that have been developed so far in the distributional compositional model. We then proceed to define a new quantitative asymmetric measure on density matrices, called p-hyponymy, which allows us to determine the strength of hyponymy in hyponym-hypernym pairs. We show how this can be lifted to the level of the sentence structures that our mathematical model of meaning supports and consider the implications of this result. We conclude with a brief discussion of how this measure can potentially be modified to account for other similar phenomena.

Contents

Chapter 1

Introduction

1.1 Background

Faithfully representing meaning in natural languages using mathematical formalism is one of the most challenging questions in linguistics and computer science. The practical gains of acquiring a deeper formal understanding of language include improvements to tasks such as machine translation, document retrieval and search optimisation, among others. On the theory front, the process of developing and exploring mathematical structures that better capture language has the potential to further our understanding of cognition and intelligence.

Computers are very good at tasks that involve words existing in a vacuum, but when these words are put together into larger grammatical units, such as sentences, they fail to produce the same high-quality results. The main problem is that sentences are not simply concatenations of words in which even the word order is irrelevant. In fact, there is a very deep connection between the meaning of a sentence, the meaning of its words, and the grammar that propagates through it in order to output a coherent whole. We humans excel at deducing meanings of sentences we have never seen before by simply drawing upon our knowledge of grammar and vocabulary. This, it turns out, is an incredibly complex task for a computer.

Two orthogonal solutions to this problem have been explored over the years. The more traditional approach [27] is built upon concepts from classical mathematical logic and adheres to the principle of compositionality [30], which tells us that the meaning of a sentence is a function of the syntactic relationships of the words comprising it. Basic grammatical types are assigned to words, which are represented as elements in an algebraic structure such as a Lambek grammar or pregroup [23, 24], and interactions between words are achieved via the binary operation with which the structure is equipped. This method, however, completely disregards the actual meanings of words.

This leads us to the second and much more recent approach, the distributional one, which is based on Firth's dictum that 'You shall know a word by the company it keeps' [8]. In this model, words are represented via vectors, usually from high-dimensional vector spaces whose basis elements correspond to relevant context features. These are normally extracted from a large body of text such a corpus. The idea behind representing words in terms of other co-occurring words lies in the assumption that meaning is based entirely on contextual co-occurrence. This approach has proved to be very useful in practical applications, but has very limited theoretical value and does not provide us with the whole picture either.

To sum up, the main difference between the compositional and the distributional models of meaning is that the former is qualitative and theoretical, while the latter is quantitative and practical. Neither of them is capable of fully capturing meaning. The quest for finding a more complete mathematical framework for the natural language problem has led in recent years to the development of a new model, unifying the above-mentioned ones. First introduced in [4] and formalised in [7], the so-called distributional compositional categorical (DisCoCat) model of meaning unites the structures used in the compositional and the distributional approach, namely pregroups and finite dimensional vector spaces, via category theory and draws inspiration from concepts, ideas, results and formalism from quantum information theory.

In this model, the meaning of a string of words is computed via a categorical morphism extracted from the grammar of this string and applied to the tensor product of the vectors corresponding to its functional words. Upon application, these so-called meaning maps give an output of the form of a vector and vectors corresponding to meanings of structures of the same grammatical type always live in the same vector space. This enables us to compare meanings above the word level using the same similarity measures as for words, such as the vector cosine. However, in many cases in language and in phenomena from cognitive linguistics, asymmetry in similarity is prevalent. Prototypicality and hyponymy, asymmetry of similarity judgments and overextension with respect to concept combination are just some of the examples where making comparisons under the assumption of symmetry fails to produce adequate results.

In this dissertation we will attempt to utilise the DisCoCat framework developed so far, including its recent extension that allows the use of vectors to be replaced by density matrices, to model asymmetry in a number of different scenarios. Work in the area has only recently began with [2, 32, 33] and there is still a lot of room for the development of various approaches to capturing hyponymy, entailment, prototypicality and many other intrinsically asymmetric phenomena.

1.2 Overview and new contributions

We begin Chapter 2 with an overview of the mathematical framework behind the distributional compositional model of meaning of [7]. We first introduce pregroup grammars and finite dimensional Hilbert spaces as standalone structures, while simultaneously outlining their use in modeling syntax and semantics, respectively. We then proceed to establish the connection between the two via the common language of category theory. We define the categories **Preg** and **FHilb**, which are both examples of the so-called compact closed categories that come equipped with a very intuitive graphical calculus, allowing us to reduce equational expressions within the categories to simple diagrammatic manipulations. We show how the structural morphisms of the categories give rise to sentence meaning maps and examine the additional structure provided by the Frobenius algebras for capturing meaning of relative clauses [34, 35].

The development of the so-called CPM construction in categories has led to recent advancements in the field whereby density matrices are used in place of vectors for word meaning [2, 32, 33]. This is achieved by passing from the category **FHilb** to CPM(**FHilb**), which is also a †-compact closed category. In Chapter 3, we outline the general framework of the CPM construction and consider how meaning maps can be interpreted in this setting and what the advantages of working with density matrices are. In brief, density matrices showcase more clearly the difference between mixing and correlation of features, similar to their role in quantum computing. In addition, they possess a richer structure that enables us to consider various measures of similarity that vectors do not admit.

We begin Chapter 4 with a brief discussion of a couple of linguistic phenomena from psychology and cognition, namely the problem of overextension with respect to concept combination [17] and asymmetry of similarity judgment [40]. While these phenomena have been around for a long time, there is still a lot of room for improvement in terms of the mathematical models used to capture them. Recent work [25] has shown some promise in applying the DisCoCat model to represent overextension via the traditional vector-based approach by taking into account the grammatical structure of the individual concepts that comprise a combined unit. Here we extend upon this work to show how density matrices can be used for the same task. We then go on to show how asymmetry in measuring the similarity between a more prototypical concept and another concept from the same category can be represented by using the intrinsic asymmetry of verbs such as *is similar to* and the compositonality of the meaning map. We come back to the same phenomenon in Chapter 5 and consider a different approach in which the need for an explicit verb is eliminated and concepts are represented via density matrices.

One of the main advantages of transitioning to the CPM(**FHilb**) category and using density matrices instead of vectors lies in the opportunity for defining various asymmetric measures on the matrices. These in turn give rise to orderings that can be utilised in tasks such as hypernym-hyponym classification. We define a new very simple and intuitive measure on density matrices, called p-hyponymy, that allows us to extract quantitative information about the relative strength of a given hyponym-hypernym bond. We show how this link manifests itself at the sentence level in a variety of different sentence structures. Finally, we briefly discuss the possibility of implementing variations of this measure to the task of modeling linguistic phenomena other than hyponymy, such as those from Chapter 4.

Chapter 2

A compositional distributional model of meaning

In their paper [7], Coecke, Sadrzadeh and Clark first establish the category-theoretic framework for a new model of meaning that aims to unite the two standard approaches to capturing the structure of natural languages - the compositional and the distributional one. At first sight the use of category theory - a field which is often affectionately referred to as ‘general abstract nonsense’ - to model a natural language may seem a bit surprising. However, its suitability for this purpose is easy to observe. It enables us to model syntax and semantics separately in two different categories which, via their shared structural identities, allow us to derive sentence meanings in a way that takes into account grammar and individual word meanings simultaneously.

In more concrete terms, we store qualitative information about word meanings in the category **FHilb** and information about syntax in the category **Preg** and achieve the transition from the latter to the former via a strongly monoidal functor **Preg** → **FHilb**.

Moreover, the structural morphisms provided by the compact closed category allow us to compute maps which can be applied to strings of words to produce their combined meaning. This can be used not only for the purposes of establishing whether or not a sentence is true or false (as in Montague semantics) but also to extract richer information about it, depending on what we aim to achieve and what data we are interested in obtaining.

This chapter serves as an introduction and overview of the categorical framework used in the distributional compositional categorical (DisCoCat) model and the applications to extracting meaning out of various structures, such as positive transitive sentences and relative clauses.

2.1 Types and distributions. Grammar and meaning.

2.1.1 Pregroup Grammars

In 1958 [22] J. Lambek introduced a syntactic calculus that formalises the grammatical structure of language and later on built upon this work by making use of the mathematical formalism of pregroups [23]. Below we give a brief overview of how pregroup grammars can be used to capture syntax and grammatical reductions. For a more detailed account of the use of types as elements of a pregroup and applications, see [21, 24].

Definition 1 (Partially ordered monoid). *A* partially ordered monoid $(P\,,\, \leq\,,\, \bullet\,,\, 1)$ *is a partially ordered set* $(P\,,\, \leq)$ *together with an associative binary operation* $\bullet : P \times P \to P$ *called* monoid multiplication*, and a unit element* 1*, and such that the multiplication is order-preserving:*

$$(p \leq q) \implies (r \bullet p \leq r \bullet q) \wedge (p \bullet r \leq q \bullet r) \qquad \forall\, p, q, r \in P.$$

Definition 2 (Pregroup algebra). *A* pregroup algebra $\left(P\,,\, \leq\,,\, \bullet\,,\, 1\,,\, (-)^l\,,\, (-)^r\right)$ *is a partially ordered monoid together with two unary operations* $(-)^l : P \to P$ *and* $(-)^r : P \to P$ *called the* left *and* right adjoint *respectively, such that for each* $p \in P$ *there exist* p^l, $p^r \in P$ *satisfying:*

$$p^l \cdot p \leq 1 \leq p \cdot p^l$$
$$p \cdot p^r \leq 1 \leq p^r \cdot p.$$

Adjoints are unique and satisfy the following properties [7]:

1. *Order-reversing:* $(p \leq q) \implies (q^r \leq p^r) \wedge (q^l \leq p^l)$;
2. *Opposite adjoints annihilate:* $(p^r)^l = p = (p^l)^r$;
3. *Self-adjoint unit and multiplication:* $1^r = 1 = 1^l$ and $((p \cdot q)^r = q^r \cdot p^r) \wedge \big((p \cdot q)^l = q^l \cdot p^l\big)$.

Definition 3 (Pregroup grammar [33]). *A pregroup grammar* $\mathcal{G} = \big(P\,,\leq\,,\,\cdot\,,\,1\,,\,(-)^l\,,\,(-)^r\big)$ *is a pregroup algebra which is freely generated over a set of basic types* $\mathcal{B}$ *that includes an end type and a type dictionary which associates pregroup elements to the vocabulary of a (natural) language.*

Note that we say that a pregroup P is *freely generated* over a set $\mathcal{B}$ to mean that all the elements of P can be formed out of the elements of $\mathcal{B}$ via zero or more applications of the monoid operation and the adjoint operators $(-)^r$ and $(-)^l$.

In the context of natural languages $\mathcal{B}$ is often taken to be the set $\mathcal{B} = \{n, s, j, \sigma\}$, where n is the type assigned to nouns by the type dictionary; j is the the type of infinitives of verbs; σ is a gluing type, and s is the type of a declarative statement. We also call the type s the *end type* of this grammar. For the rest of this dissertation the pregroup grammar $\mathcal{G}$ will be understood to mean exactly the grammar that is freely generated over the set $\mathcal{B}$ above.

The following remarks, terminology and conventions will become useful later on:

- The **grammatical type** of a string of words is the concatenation (i.e. the monoidal multiplication) of the types of the individual words in the string in the order in which they appear. This is an element of $\mathcal{G}$.
- We can often simplify a grammatical type by using the properties of adjoints and the associativity of the monoidal operation. We call such a simplification a **reduction**. When a string $x \in \mathcal{G}$ can be reduced to some other string $y \in \mathcal{G}$ we write '$x \leq y$' or '$x \to y$'. I will use these interchangeably. This notion will be made more precise once compact closed categories are introduced.
- A **well-typed** or **grammatical sentence** is a sentence whose grammatical type reduces to the end type s.
- A **well-typed noun phrase** is is a phrase or a sentence whose grammatical type reduces to the basic type n.

The most important functional words that we will need here are transitive and intransitive verbs, adjectives and various pronouns. Thus, we summarise their types below, following the conventions of [34, 35].

functional word	**type**
transitive verb	$n^r s n^l$
intransitive verb	$n^r s$
adjective	$n n^l$
subject relative pronoun (who, which, that)	$n^r n s^l n$
object relative pronoun (whom, which, that)	$n^r n n^{ll} s^l$
subject possessive pronoun (whose)	$n^r n s^l n n^l$
object possessive pronoun (whose)	$n^r n n^{ll} s^l n^l$

For example, the type of a transitive verb is meant to reflect the fact that it is a functional word that takes a noun (its subject) on the left and another noun (its object) on the right in order to produce a declarative statement.

To see these types and type reductions in action, consider the following examples.

(1) postmodern paintings $(nn^l)n$

reduction: $(nn^l)n \leq n(n^l n) \leq n$

Well-typed noun phrase.

(2) Mary likes postmodern paintings. $n(n^r s n^l)(nn^l)n$

reduction: $n(n^r s n^l)(nn^l)n \leq (nn^r)s(n^l n)(n^l n) \leq s$

Well-typed sentence.

(3) Mary likes jumps. $n(n^r s n^l)(n^r s)$

reduction: $n(n^r s n^l)(n^r s) \leq (nn^r)s(nn^r s) \leq snn^r s$

Grammatical type cannot be reduced further. Not a well-typed sentence.

(4) John dislikes the postmodern paintings that Mary buys.

$$n(n^r s n^l)(nn^l)n(n^r n n^{ll} s^l n)n(n^r s n^l)$$

$$\begin{aligned} \textit{reduction:}\;\; n(n^r s n^l)(nn^l)n(n^r n n^{ll} s^l)n(n^r s n^l) &\leq (nn^r)s(n^l n)(n^l n)n^r n n^{ll} s^l (nn^r) s n^l \\ &\leq s n^l (nn^r) n n^{ll} (s^l s) n^l \\ &\leq s(n^l n)(n^{ll} n^l) \\ &\leq s \end{aligned}$$

Well-typed sentence.

2.1.2 Finite dimensional Hilbert spaces

In contrast to the pregroup grammar formalism that enables us to capture the syntactic structure of sentences, the idea behind the *distributional model of meaning*, first introduced by Firth in 1957 [9] and formalised for practical applications in the last couple of decades, is that word meanings can be modeled solely on the basis of the contexts in which these words appear.

In this model words, regardless of their grammatical role, occupy highly dimensional vector spaces with orthonormal basis vectors known as target words or *context words* and which can be all or a subset of lemmatised words extracted from a corpus, e.g. The British National Corpus or ukWaC. The entries of the word meaning vectors then correspond to the number of times that the word in question has appeared in the corpus in a window of n words of each corresponding context word, where n can be taken to be as small as 1, but is normally as high as about 5.

The mathematical structure that encapsulates this formalism is that of finite-dimensional real vector spaces. Here we will be more general and instead of finite-dimensional vector spaces over $\mathbb{R}$ we will consider the category of finite-dimensional Hilbert spaces and bounded linear maps, of which the former is simply a special case.

Vector space terminology and notation.

Unit vectors in a Hilbert space V will be written interchangeably as either $\overrightarrow{v} \in V$ or $|v\rangle$ where $|\bullet\rangle$ is called the Dirac *ket* and can also be treated as an operator $|\bullet\rangle : \mathbb{C} \to V$. The Dirac *bra* is the dual operator of the ket and is given by $\langle\bullet| : V \to \mathbb{C}$. We write $\langle v|$, where $\overrightarrow{v} \in V$, to represent the *effect* corresponding to the *state* $|v\rangle$. The effect is the *Hermitian conjugate* of the state. Note that for our purposes we will be using real Hilbert spaces and thus will be able to think of $|v\rangle$ as simply being a column vector and $\langle v|$ as being its transpose, i.e. the corresponding row vector.

Together a bra $\langle v|$ and a ket $|w\rangle$ form a *bracket* $\langle v\,|\,w\rangle$, which is in fact exactly the *inner product* of the vectors $\overrightarrow{v}$ and $\overrightarrow{w}$. This will be defined more rigorously later.

We will also need the *outer product* of $\langle v|$ and $|w\rangle$ for $v \in V$ and $w \in W$, $|w\rangle\langle v|$, defined via its action of states $|u\rangle \in V$ as:

$$(|w\rangle\langle v|)\,(|u\rangle) = |w\rangle\langle v\,|\,u\rangle.$$

In applications, where the underlying field will always be assumed to be $\mathbb{R}$, the outer product will simply be the dot product of a column vector $|w\rangle$ and a row vector $\langle v|$, which results in a matrix.

Finally, an *operator* on vector spaces is defined as follows.

Definition 4. *If V and W are two Hilbert spaces, an* operator *is a map of the form $\Phi : V \to W$ that can be expressed as:*

$$\Phi = \sum_{ij} \alpha_{ij}\, |w_j\rangle\langle v_i| \quad \textit{for} \quad \alpha_{ij} = \langle w_j|\Phi|v_i\rangle,$$

where $\{|v_i\rangle\}$ is a basis for V and $\{|w_j\rangle\}$ is a basis for W.

2.2 Category theoretic and graphical calculus framework for DisCoCat

The common structural framework occupied by the otherwise orthogonal models of meaning provided by the type-theoretic and the distributional models is that of compact closed categories. Before going into more detail about how this is achieved, we give a brief introduction to the basics of category theory that will be needed for our purposes. For more background on the vast and increasingly important subject that is category theory, we refer the reader to some of the many good sources on the topic [3, 6, 26].

2.2.1 What is a category?

Definition 5 (Category)**.** *A category $\mathcal{C}$ consist of the following:*

- *A collection of* objects *$Ob(\mathcal{C})$;*
- *A collection of* morphisms *$Ar(\mathcal{C})$ such that for each pair of objects $A, B \in Ob(\mathcal{C})$ there is a set of morphisms $\mathcal{C}(A, B) = \{f \in Ar(\mathcal{C}) \mid f : A \to B\} \subseteq Ar(\mathcal{C})$, called a* hom-set*;*
- *For any pair of morphisms $f \in \mathcal{C}(A, B)$ and $g \in \mathcal{C}(B, C)$, a composite morphism $g \circ f \in \mathcal{C}(A, C)$, satisfying the following axioms:*

  - associativity: $\forall f \in \mathcal{C}(A, B)$, $\forall g \in \mathcal{C}(B, C)$, $\forall h \in \mathcal{C}(C, D)$,

  $$h \circ (g \circ f) = (h \circ g) \circ f$$

  - identity: $\forall\, A \in Ob(\mathcal{C})\ \exists!\ id_A \in \mathcal{C}(A, A)$ *s.t.* $\forall\, f \in \mathcal{C}(A, B)$ *and for any* $B \in Ob(\mathcal{C})$,

  $$f = f \circ id_A = id_B \circ f.$$

2.2.2 Monoidal categories

We will now define a particular subclass of categories that will prove to be especially useful for our purposes. Before that, we will need the notion of a functor, as well as a few types of functors that will be applicable later on.

Definition 6 (Functor [1])**.** *Let $\mathcal{C}$ and $\mathcal{D}$ be two categories. A* functor *$\mathcal{F} : \mathcal{C} \to \mathcal{D}$ is given by:*

- *A mapping on objects: $\mathcal{F} : Ob(\mathcal{C}) \to Ob(\mathcal{D})$ by $A \mapsto \mathcal{F}A$;*

- *A mapping on morphisms:* $\mathcal{F} : Ar(\mathcal{C}) \to Ar(\mathcal{D})$ *by* $(f \in \mathcal{C}(A,B)) \mapsto (\mathcal{F}f \in \mathcal{D}(\mathcal{F}A, \mathcal{F}B))$*, which preserves identities and compositions:*

$$\mathcal{F}_{id_A} = id_{\mathcal{F}A} \quad (\forall\, A \in Ob(\mathcal{C})) \quad \textit{and} \quad \mathcal{F}(g \circ f) = \mathcal{F}g \circ \mathcal{F}f \quad (\forall\, f, g \in Ar(\mathcal{C})).$$

Definition 7 (Monoidal functor; Strongly monoidal functor [1])**.** *If* $\mathcal{C}$ *and* $\mathcal{D}$ *are two monoidal categories and* $\mathcal{F} : \mathcal{C} \to \mathcal{D}$*, we say that* $\mathcal{F}$ *is a* monoidal functor *to mean that, in addition to* $\mathcal{F}$ *being a functor, we also have a natural transformation such that* $\forall A, B \in Ob(\mathcal{C})$, $\mathcal{F}(A) \otimes \mathcal{F}(B) \to \mathcal{F}(A \otimes B)$*, and a morphism* $I \to \mathcal{F}I$*, where* $I \in Ob(\mathcal{C})$ *is the unit object of* $\mathcal{C}$ *. Whenever these are both invertible we say that* $\mathcal{F}$ *is a* strongly monoidal functor*.*

Definition 8 (Dagger functor)**.** *A dagger functor is a functor* $\dagger : \mathcal{C} \to \mathcal{C}^{op}$ *such that for all* $\varphi \in Ar(\mathcal{C})$ *we have* $\left(\varphi^{\dagger}\right)^{\dagger} = \varphi$*.*

Definition 9 (Monoidal category)**.** *A monoidal category* $(\mathcal{C}\,,\,\otimes\,,\,I\,,\,a\,,\,l\,,\,r)$ *consists of:*

- *A category* $\mathcal{C}$*;*
- *A functor* $\otimes : \mathcal{C} \times \mathcal{C} \to \mathcal{C}$ *called a* tensor *such that it acts of objects by:*

$$(A, B) \mapsto A \otimes B \in Ob(\mathcal{C}),$$

and on morphisms by:

$$(f \in \mathcal{C}(A,B),\, g \in \mathcal{C}(C,D)) \mapsto f \otimes g \in \mathcal{C}(A \otimes B, C \otimes D)\,.$$

Moreover, this functor is bifunctorial, meaning that for all $f, g, h, k \in Ar(\mathcal{C})$ *we have:*

$$(f \otimes g) \circ (h \otimes k) = (f \circ h) \otimes (g \circ k).$$

- *A distinguished object* $I \in Ob(\mathcal{C})$ *called* unit object*.*
- *Natural isomorphisms* a*,* l*,* r *whose components are given by:*

$$(\forall\, A, B, C \in Ob(\mathcal{C})) \quad a_{A,B,C} : A \otimes (B \otimes C) \xrightarrow{\cong} (A \otimes B) \otimes C$$

$$(\forall A \in Ob(\mathcal{C})) \quad l_A : I \otimes A \xrightarrow{\cong} A \textit{ and } r_A : A \otimes I \xrightarrow{\cong} A \quad \textit{, with } r_I = l_I : I \otimes I \xrightarrow{\cong} I.$$

Moreover, these natural isomorphisms have to satisfy certain coherence conditions which ensure that all the relevant diagrams commute.

A *symmetric monoidal category* is a monoidal category equipped with a swap map $\sigma : A \otimes B \xrightarrow{\cong} B \otimes A$, for any pair of objects A and B is the category.

Finally, a category is a *dagger category* if it is equipped with a dagger functor. Note that dagger categories have a richer structure and satisfy additional criteria not mentioned here as these will not be of direct relevance to the current discussion. Thus, we omit the details and refer the reader to [37].

The main reasons why monoidal categories are so useful for the purposes of modeling meaning lie in the existence of the monoidal tensor $\otimes$ and the identity object.

The monoidal tensor allows us to consider situations where several objects (words) need to be looked at at the same time as a sequence, or when several processes (morphisms) take place simultaneously. Loosely speaking, one can think of the tensor of two objects $A, B \in Ob(\mathcal{C})$, $A \otimes B$, as being 'object A *and* object B' and the tensor of two morphisms $f, g \in Ar(\mathcal{C})$ as being process f and process g occurring *at the same time*. The latter complements the sequential composition of processes provided by the categorical morphism composition operation which tells us that we can interpret $f \circ g$ as 'process f happens *after* process g'. For a more detailed explanation and notes on how this applies in quantum computing as well, see [6]. The availability of both parallel and sequential processes in monoidal categories is what makes them a good candidate for a framework in which sentence meaning can be interpreted, as we will see shortly.

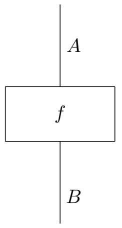
The identity element $I \in Ob(\mathcal{C})$ allows us to think down to the level of the elements of which the objects of the category are built, while at the same time retaining the generality provided by the categorical formalism. This is because the properties of the identity object imply the existence of a bijective correspondence between the actual elements of $A \in Ob(\mathcal{C})$ and morphisms of the type $a : I \to A$, where $a \in \mathcal{C}(I, A)$. This correspondence will allow us to think of words as being linear maps in a vector space and at the same time its elements.

Graphical Calculus for monoidal categories

Another significant advantage of monoidal categories is that they are complete with respect to a very intuitive graphical calculus. By completeness we mean that any statement about the equality of morphisms in a monoidal category can be derived from the categorical axioms if and only if the same statement can be obtained via an admissible sequence of manipulations of diagrams expressible in the monoidal graphical calculus.

The origins of this useful graphical language date back to Roger Penrose [31], who first proposed representing morphisms (processes) as boxes and objects (input and output types) as wires. For more details about the origins and development of the graphical calculus, its applications in categorical quantum computing, and a proof of the above statement, see [5, 38]. We will now only present the most basic diagrams and supplement these with other constructions as necessary later. The main building blocks of the diagrammatic calculus are boxes of various shapes and wires, which also give rise to the name *string diagrams*.

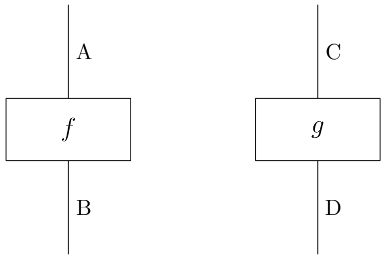
If $A, B \in Ob(\mathcal{C})$ and $f \in \mathcal{C}(A, B)$ is a morphism then we represent it as a box with input wire labeled A and output wire labeled B. We adopt the convention here that *information flows from top to bottom.*

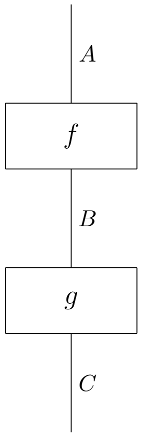


For any $A, B, C, D \in Ob(\mathcal{C})$ and $f \in \mathcal{C}(A, B)$, $g \in \mathcal{C}(C, D)$ the parallel morphism composition, i.e. the tensor of morphisms

$$f \otimes g \in \mathcal{C}(A \otimes C, B \otimes D)$$

is depicted by simply placing the corresponding boxes next to each other.



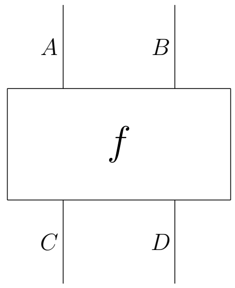


For any $A, B, C \in Ob(\mathcal{C})$ and $f \in \mathcal{C}(A, B)$, $g \in \mathcal{C}(B, C)$ the sequential composition

$$g \circ f \in \mathcal{C}(A, C)$$

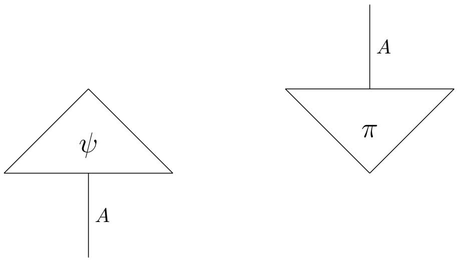
is represented by simply stacking the boxes on top of each other, with the flow of information from one process of the other carried via the wire of their common type B.

Note that if a morphism has several inputs and/or outputs then we represent these as different wires into/out of the appropriate morphism box. For example, $f \in \mathcal{C}(A \otimes B, C \otimes D)$ as a single morphism takes the form

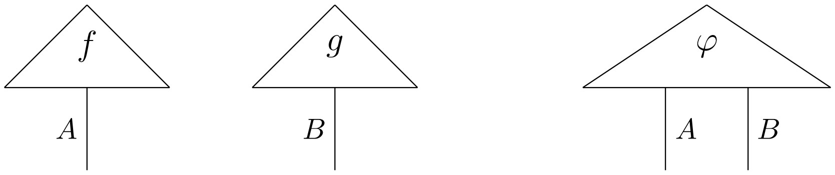


The identity object $I \in Ob(\mathcal{C})$ is depicted by an empty box and for each $A \in Ob(\mathcal{C})$ the identity morphism $1_A : A \to A$ takes the form of a single wire of type A.

Whenever we have a morphism with no inputs or outputs, i.e. when the input or outputs are of type I, we depict these as up and down triangles respectively. These are normally denoted by $\varphi : I \to A$ and $\pi : A \to I$ and are called *states* and *effects*, following the quantum terminology.

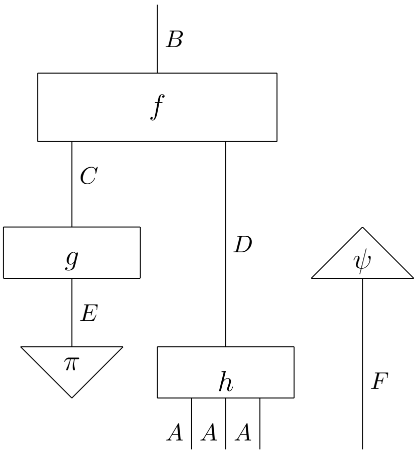


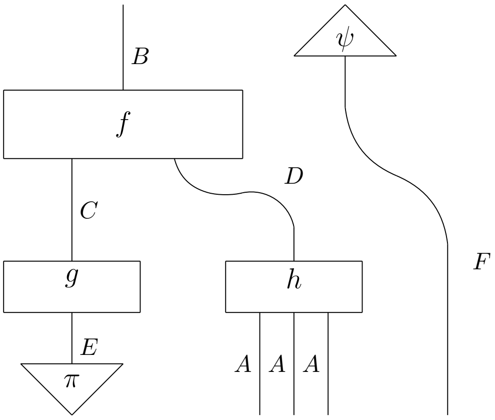
[18] If A and B are two objects in our monoidal category then we call a morphism $f : I \to A \otimes B$ a *joint state* of A and B. A joint state is said to be a *product state* or *separable* if it has the form $\varphi : I \to I \otimes I \xrightarrow{f \otimes g} A \otimes B$, where $f : I \to A$ and $g : I \to B$. Joint states which are not product states are called *entangled* states. The diagram on the left hand side depicts a product state and the one on the right is an entangled state.



Entangled states represent tensors $A \otimes B$ which cannot be decomposed into A and B as there is some kind of intrinsic interconnectedness between the two. For example, we could have an element $\omega \overrightarrow{a} \otimes \overrightarrow{b} \in A \otimes B$ which cannot be separated into $\rho \overrightarrow{a} \in A$ and $\sigma \overrightarrow{b} \in B$. This is very handy when depicting functional words like verbs since in such cases we want to be able to represent the inseparability of the constituents. This is essentially what allows us to put together a sentence in which the words have a way of interacting with each other, rather than simply existing next to one another in a string. This will become more clear when we define functional words in tensor spaces explicitly and start performing calculations.

More complicated processes are depicted diagrammatically by combining these building blocks. Note that the topology of the diagrams, i.e. the relative positions of the boxes and wires is immaterial as the only thing of importance is how these are connected. For more on this refer to [5]. For example, the following two diagrams, and their corresponding morphisms, are in fact equal:





$(\pi \otimes h \otimes 1_F) \circ (g \otimes 1_D \otimes \psi) \circ f$

$(\pi \otimes 1_A \otimes 1_A \otimes 1_A \otimes 1_F) \circ (g \otimes h \otimes 1_F) \circ (f \otimes 1_F) \circ \psi$

2.2.3 Compact closed categories

Now we only need the morphisms for interpreting the grammatical reductions defined in the framework of the pregroup grammars. These are provided by compact closed categories.

Definition 10 (Compact closed category)**.** *A* compact closed category *$\mathcal{C}$ is a monoidal category in which for any object $A \in Ob(\mathcal{C})$ there exists a pair of objects A^r and A^l also in $Ob(\mathcal{C})$, called the* right *and* left adjoint *of A, and corresponding structural morphisms ε^r_A, ε^l_A. η^r_A, η^l_A given by:*

$$A \otimes A^r \xrightarrow{\varepsilon^r_A} I \qquad A^l \otimes A \xrightarrow{\varepsilon^l_A} I \qquad I \xrightarrow{\eta^r_A} A^r \otimes A \qquad I \xrightarrow{\eta^l_A} A \otimes A^l$$

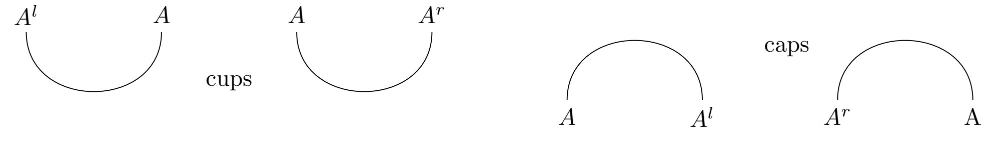
The first two of these are also known as ***cancellations*** *and the second pair as* ***generations****. These satisfy the following equations, known as the* yanking equations*:*

$$(1_A \otimes \varepsilon^l_A) \circ (\eta^l_A \otimes 1_A) = 1_A$$
$$(\varepsilon^r_A \otimes 1_A) \circ (1A \otimes \eta^r_A) = 1_A$$
$$(\varepsilon^l_A \otimes 1_A) \circ (1_{A^l} \otimes \eta^l_A) = 1_{A^l}$$
$$(1_{A^r} \otimes \varepsilon_{A^r}) \circ (\eta^r_A \otimes 1_{A^r}) = 1_{A^r}$$

Graphical calculus for compact closed categories

The structural morphisms are depicted as caps and cups in the languages of the graphical calculus accompanying monoidal categories. We will use the following conventions:

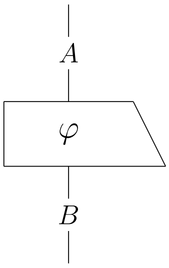
$\varepsilon^l_A : A^l \otimes A \to I$ $\qquad$ $\varepsilon^r_A : A \otimes A^r \to I$



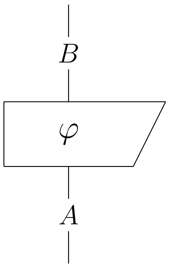
$\eta^l_A : I \to A \otimes A^l$ $\qquad$ $\eta^r_A : I \to A^r \otimes A$

Graphical calculus for †-compact closed categories

It will be easier to adopt the convention of drawing morphisms as asymmetric boxes when they inhabit a dagger compact closed category. Note that this convention will not be applied to states and effects which will still be depicted as triangles. So, for a general morphism $\varphi : A \to B$ we have:



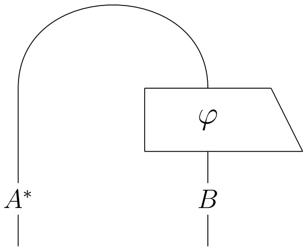
Then we depict application of the † functor to φ as a reflection of the box along the horizontal axis, i.e $\varphi^\dagger : B \to A$ is depicted as:

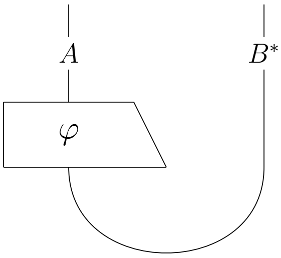


Definition 11 (Name, Coname [18])**.** *Let $\mathcal{C}$ be a †-compact closed category. We define the* name *and* coname *of a morphism $\varphi \in \mathcal{C}(A, B)$ to be (respectively) the morphisms:*

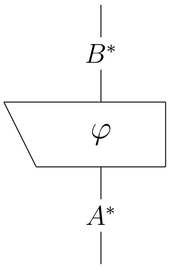
$$\ulcorner \varphi \urcorner : I \to A^* \otimes B \qquad \llcorner \varphi \lrcorner : A \otimes B^* \to I$$

$$\ulcorner \varphi \urcorner = (id_{A^*} \otimes \varphi) \circ \eta_A \qquad \llcorner \varphi \lrcorner = \varepsilon_B \circ (\varphi \otimes id_{B^*})$$

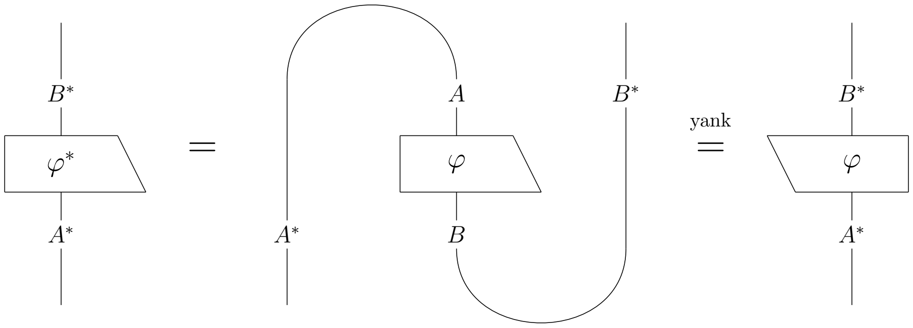




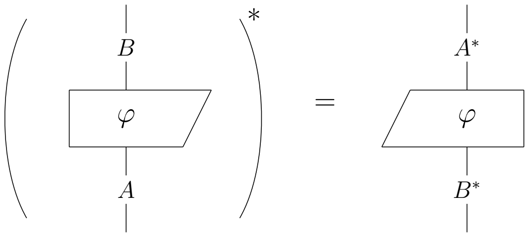
Definition 12 (Dual morphism [18])**.** *Define the* dual *of a morphism $\varphi : A \to B$ to be $\varphi^* : B^* \to A^*$, given by:*



which is obtained via



Finally, we define $\varphi_* = \left(\varphi^\dagger\right)^*$ to be:



Note that applying † to the ε and η maps has the effect of reflecting them along the horizontal axis.

2.3 Uniting syntax and semantics via a compact closed category

We are finally in the position to describe the structures that we already defined as the containers for grammar and meaning in the language of category theory.

2.3.1 Pregroup grammars as compact closed categories

A pregroup grammar $\mathcal{G} = \Big(P\,,\, \leq\,,\, \boldsymbol{\cdot}\,,\, 1\; (\boldsymbol{\cdot})^r\,,\, (\boldsymbol{\cdot})^l\Big)$ is a compact closed category $\mathcal{C} = \mathbf{Preg}$:

- The *objects* are the elements of the underlying set P.
- The *morphisms*, denoted by '$\rightarrow$' or '$\leq$', correspond to the partial order $\leq$ between the elements of P in the sense that there is a morphism between $p, q \in Ob(\mathcal{C})$ iff $p \leq q$ as elements of P. Note that if there exists a morphism between any pair of objects then it is necessarily unique. In other words, between any two objects of the category there is either one morphism or none.
- The existence of *composite morphisms* follows from the transitivity of the partial order $\leq$ of P:
$$(p \leq q) \wedge (q \leq r) \implies (p \leq r) \qquad \forall\, p, q, r \in P.$$
We can also express this by saying that '$p \rightarrow q$' and '$q \rightarrow r$' implies '$p \rightarrow r$' for any $p, q, r \in Ob(\mathcal{C})$.
- The existence of an *identity morphism* on any object $p \in Ob(\mathcal{C})$ follows from the reflexivity of the partial order:
$$p \leq p \qquad \forall\, p \in P.$$
We also write '$p \rightarrow p$'.
- The *monoidal tensor* $\otimes$ of $\mathcal{C}$ is the monoidal multiplication $\boldsymbol{\cdot} : P \times P \longrightarrow P$ of $\mathcal{G}$.

  - The tensor on objects $p \otimes q \;\; (p, q \in Ob(\mathcal{C}))$ is given by $p \boldsymbol{\cdot} q$, which we simply write as pq.

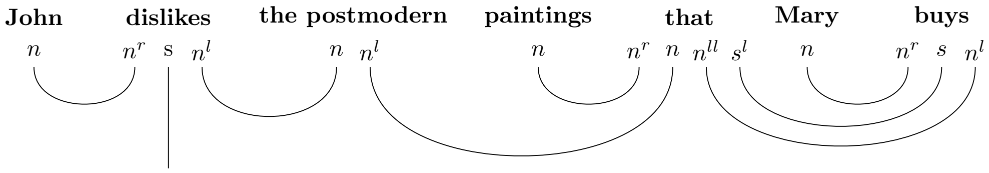
  - The tensor on morphisms follows from the transitivity of the partial order on P and the order-preserving property of the monoidal multiplication as follows:
$$[((p \leq q) \implies pr \leq qr) \wedge ((r \leq s) \implies qr \leq qs)] \implies pr \leq qr \leq qs \implies pr \leq qs.$$

- The *left* and *right adjoints* of $p \in Ob(\mathcal{C})$ are simply the left and right adjoints of the element $p \in P$, which exists by definition of $\mathcal{G}$.
- The structure-preserving ε and η maps are given by:

$$\varepsilon^l = p^l \cdot p \to 1 \qquad\qquad \eta^l = 1 \to p \cdot p^l$$
$$\varepsilon^r = p \cdot p^r \to 1 \qquad\qquad \eta^r = 1 \to p^r \cdot p$$

Note that we may alternatively write '$\leq$' instead of '$\to$', e.g. $\varepsilon^l = [p^l \cdot p \leq 1]$. One can easily verify that these satisfy the compact closure axioms.

Let $\mathcal{B} = \{n, s, \sigma, j\}$ be the set of basic grammatical types that we had before. We will denote by $\mathbf{Preg}_{\mathcal{B}}$ the compact closed category corresponding to the pregroup grammar $\mathcal{G}$ used to model the grammar and grammatical reductions of strings of words. As a compact closed category, $\mathbf{Preg}_{\mathcal{B}}$ is accompanied by a graphical calculus. This comes very handy as the graphical depictions of the ε and η maps greatly simplify grammatical reductions. To see this, consider the reduction of one of our previous examples done via a string diagram:



What this diagrams tells us is that the output type of the combined morphisms is simply s, as this is the type of the only outgoing wire. Thus, the type of the sentence is s, as expected.

2.3.2 Finite dimensional Hilbert spaces as compact closed categories

The category in which we will model meaning is **FHilb**. Note that in the literature, meaning is often modeled in the category of finite-dimensional (real) vector spaces and linear maps **FVect**. In practice, which of these is used for the purposes of any of the applications mentioned in the present work makes no difference, as in either case we restrict our attention to real vector spaces and a very narrow set of morphisms. The advantage of working with **FHilb** is that it allows for more structure and is equipped with a canonical inner product that gives rise to adjoints. It is is a motivating example for the class of dagger compact closed categories used in the CPM construction, which will be elaborated on in the next chapter.

Let H be an arbitrary Hilbert space. Then H has an inner product $\langle \cdot \,|\, \cdot \rangle_H : H \times H \to \mathbb{C}$, which is:

- antilinear in the first argument: $\langle \lambda f + g \,|\, h \rangle_H = \lambda \langle f \,|\, g \rangle_H + \overline{\lambda}\, \langle g \,|\, h \rangle_H \quad (\lambda \in \mathbf{C})$;
- linear in the second argument: $\langle f \,|\, \lambda g + h \rangle_H = \lambda \langle f \,|\, g \rangle_H + \lambda \langle f \,|\, h \rangle_H$;
- conjugate-symmetric: $\langle f \,|\, g \rangle_H = \overline{\langle g \,|\, f \rangle}_H$;
- positive semi-definite: $\langle f \,|\, f \rangle_H \geq 0$.

Note that when there is no ambiguity, we will simply write $\langle \cdot \,|\, \cdot \rangle$ instead of $\langle \cdot \,|\, \cdot \rangle_H$. The adjoint of a liner maps is now defined via the canonical inner product.

Definition 13 (Adjoint of linear map). *If V and W are Hilbert spaces and $\varphi : V \to W$ is a linear map between them then its* adjoint *is defined to be the unique linear map $\varphi^\dagger : W \to V$ such that $\forall f \in V$ and $g \in W$, we have:*

$$\langle \varphi f \,|\, g \rangle_W = \langle f \,|\, \varphi^\dagger g \rangle_V .$$

This defines a natural dagger functor on **FHilb**.

Definition 14 (Adjunctor functor). *The adjunctor functor $\dagger : \mathbf{FHilb} \to \mathbf{FHilb}^{op}$ is a functor which preserves objects and takes morphisms $\varphi \in Ar(\mathbf{FHilb})$ to their adjoints $\varphi^\dagger \in Ar(\mathbf{FHilb}^{op})$. This is a dagger functor, i.e. $\left(\varphi^\dagger\right)^\dagger = \varphi$.*

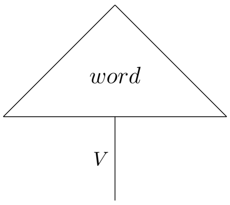
We can now formally define the category **FHilb** as the †-compact closed category with:

- *Objects*: finite dimensional Hilbert spaces over $\mathbb{C}$.
- *Morphisms*: bounded $\mathbb{C}$-linear maps.
- *Morphism composition* is provided by the closure under composition of $\mathbb{C}$-linear maps.
- The *unit object* is the field itself.
- The *monoidal tensor* is the vector tensor product $\otimes$.
- The *left* V^l and *right* V^r *adjoints* of $V \in Ob(\mathbf{FHilb})$ are both given by the dual vector space V^* of V. Note that by fixing a basis $\{\overrightarrow{e_i}\}_i$ for V we have that $V^* \cong V$. Thus, from now on we will adopt the convention of writing V to mean any of these: V, V^r, V^l, V^*.
- The *structure-preserving* maps ε and η are given by:

$$\varepsilon^r_V = \varepsilon^l_V = \varepsilon_V : V \otimes V \longrightarrow \mathbb{R} \quad by \quad \sum_{ij} \alpha_{ij}\, \overrightarrow{e_i} \otimes \overrightarrow{e_j} \mapsto \sum_{ij} \alpha_{ij} \langle \overrightarrow{e_i} \mid \overrightarrow{e_j} \rangle$$
$$\eta^r_V = \eta^l_V = \eta_V : \mathbb{R} \longrightarrow V \otimes V \quad by \quad 1 \mapsto \sum_i \overrightarrow{e_i} \otimes \overrightarrow{e_i}$$

  where $\{\overrightarrow{e_i}\}_i$ is a basis for V and $\{1\}$ is a basis for $\mathbb{R}$. As we will only be interested in the restriction to real Hilbert spaces, this definition suffices.

For the time being, the most important type of morphisms in **FHilb** for us will be the states. Recall that a state is a morphism from the unit object to another object. In this case, it is a linear map of the form $\mathbb{R} \to V$ where V is a (real Hilbert) vector space. These are in a one-to-one correspondence with elements of the space $v \in V$, which can be established by considering the image of $1 \in \mathbb{R}$. Thus, we will write $v : \mathbb{R} \to V$ to mean the morphism that sends $1 \mapsto v$. This allows us the flexibility of being able to think of states as morphisms and elements (of the vector space) at the same time. More concretely, it allows us to consider individual words in a sentence as vectors even though in reality they are linear maps. In other words, if we are working in some vector space V and want to depict a word that lives in this space in the form of a vector $\overrightarrow{word} \in V$, we will draw:



Note that $|v\rangle|w\rangle$ is often used as a shorthand for $\overrightarrow{v} \otimes \overrightarrow{w}$ and we will use these notations interchangeably.
Note also that $|v\rangle\langle w| \cong |v\rangle|w\rangle$ by writing the transpose of each row of the matrix $|v\rangle\langle w|$ one after the other in a single column vector.

2.3.3 Meanings of sentences

Recall that we defined a way in which we can transition between two monoidal categories $\mathcal{C}$ and $\mathcal{D}$ called a *functor*, and in particular we had a subclass of functors called *strongly monoidal functors.* A very useful property of strongly monoidal functors applied to compact closed categories is that they preserve the compact closure structure in the following sense:

$$\mathcal{F}(A^r) = \mathcal{F}(A)^r \text{ and } \mathcal{F}(A^l) = \mathcal{F}(A)^l \quad \forall A \in Ob(\mathcal{C}).$$

The transition between the category $\mathbf{Preg}_{\mathcal{B}}$ of grammar and the category **FHilb** of word meaning is achieved via a strongly monoidal functor $\mathcal{F} : \mathbf{Preg}_{\mathcal{B}} \longrightarrow \mathbf{FHilb}$. Note that it suffices to consider the action of this functor on the elements of the generating set $\mathcal{B}$ and the morphisms between elements of this set. We have:

$$\mathcal{F} : Ob(\mathbf{Preg}_{\mathcal{B}}) \longrightarrow Ob(\mathbf{FHilb}) \text{ by } x \mapsto \begin{cases} N & \text{if } x \in \{n, \sigma, j\} \\ S & \text{if } x := s \\ I = \mathbb{R} \text{ or } \mathbb{C} & \text{if } x := 1 \end{cases}$$

Here N is taken to be the vector space containing the nouns and spanned by the appropriate basis context vectors and S is the vector space that is meant to contain the meanings of sentences. Note that in practice we will sometimes take subspaces of N for the various nouns we consider, e.g. for objects and subjects of sentences, and also that there is no fixed vector space that gets assigned to S by default. We may have one or two-dimensional sentence space or even $S = N \otimes N$ depending on the problem at hand. As these will depend on specific applications, we will define them accordingly whenever necessary later in this thesis. A brief account of this is provided in the last section of this chapter.

$\mathcal{F} : Ar(\mathbf{Preg}_{\mathcal{B}}) \longrightarrow Ar(\mathbf{FHilb})$ maps partial orders between elements of $\mathcal{B}$ to linear maps between the appropriate vector spaces.

Some useful properties of the functor include:

- $\mathcal{F}(x^r) = \mathcal{F}(x) = \mathcal{F}(x^l)$ for any element x in the pregroup grammar. This follows from the fact that for finite dimensional vector spaces we have $V \cong V^*$ and hence $\mathcal{F}(V^*) = \mathcal{F}(V)$.
- $\mathcal{F}(X^{rr}) = \mathcal{F}(x) = \mathcal{F}(x^{ll})$ for any x in the grammar. This follows from $V^{**} \cong V$, and thus $\mathcal{F}(V^{**}) = \mathcal{F}(V)$.
- Functoriality tells us that if $x = s_1 \dots s_n$ is any string, i.e. element, in the pregroup grammar, then $\mathcal{F}(x) = \mathcal{F}(s_1) \otimes \dots \otimes \mathcal{F}(s_n)$.
- Preservation of the compact closure maps ε and η.

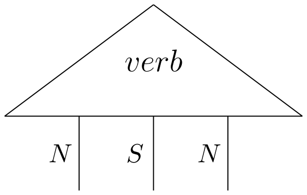
For example, we have that:

$$\begin{aligned} \mathcal{F}(n^r s n^l) = \mathcal{F}(n^r \otimes s \otimes n^l) &= \mathcal{F}(n^r) \otimes \mathcal{F}(s) \otimes \mathcal{F}(n^l) \\ &= \mathcal{F}(n) \otimes \mathcal{F}(s) \otimes \mathcal{F}(n) \\ &= N \otimes S \otimes N, \end{aligned}$$

and $n^r s n^l$ is the type of a transitive verb, so we conclude that the meanings of transitive verbs live in the tensor space $N \otimes S \otimes N$. In other words, we may represent a transitive verb as:

$$\overline{verb} = \sum_{ijk} C_{ijk}^{verb} \overrightarrow{e_i} \otimes \overrightarrow{s_j} \otimes \overrightarrow{e_k},$$

where $\{e_i\}_i$ is a basis for N and $\{s_j\}_j$ for S and the coefficients C_{ijk}^{verb} come from the underlying field $\mathbb{R}$. Diagrammatically, a verb looks like this:



Similarly, for adjectives we get $\mathcal{F}(nn^l) = \mathcal{F}(n \otimes n^l) = \mathcal{F}(n) \otimes \mathcal{F}(n^l) = N \otimes N$, and hence we can represent these as:

$$\overline{adjective} = \sum_{ij} C_{ij}^{adj} \overrightarrow{e_i} \otimes \overrightarrow{e_j}.$$

Finally, we define the meaning of a string of words as follows. Suppose that we have a string of words (not necessarily a well-typed sentence) $s = w_1 \dots w_n$ and let the type of word w_i be t_i. These types, as well as the composite type of s, $t_1 \dots t_n$, are objects in the category $\mathbf{Preg}_{\mathcal{B}}$. Now suppose that we have a type reduction $t_1 \dots t_n \xrightarrow{r} x$ for some type x that cannot be reduced any further but need not be a basic type. This reduction r is tensor and/or composition of structural morphisms of the category $\mathbf{Preg}_{\mathcal{B}}$ and can therefore be translated into the corresponding morphisms in **FHilb** via the strongly monoidal functor $\mathcal{F}$. It is exactly this translation that allows us to make the transition between grammar and meaning. More precisely, we define:

Definition 15. *The from-the-meaning-of-words-to-meaning-of-sentences map, or simply* meaning map *for a string of words $s = w_1 \dots w_n$ with grammatical reduction r is given by:*

$$\mathcal{F}(r)(\overrightarrow{w_1} \otimes \ldots \otimes \overrightarrow{w_n}).$$

For example, consider the sentence **Carnivorous animals eat meat.** The word types are, in the order in which they appear, nn^l, n, n^rsn^l, n. The types reduction r is given by:

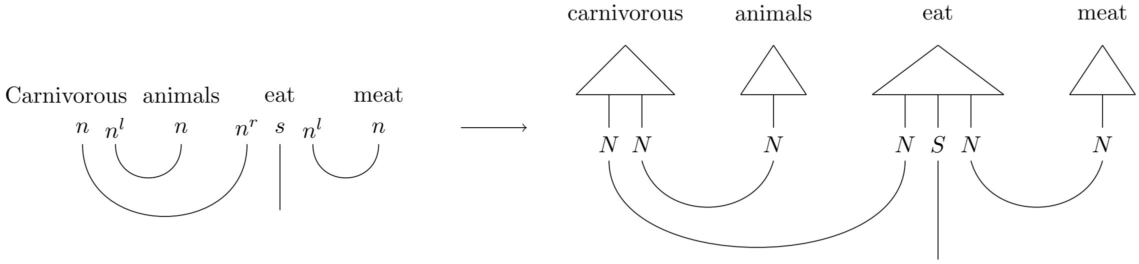
$$(nn^l)n(n^rsn^l)n = n(n^ln)n^rs(n^ln) \xrightarrow{1_n\otimes\varepsilon^l_n\otimes 1_n\otimes 1_s\otimes\varepsilon^l_n} (nn^r)s \xrightarrow{\varepsilon^r_n\otimes 1_s} s,$$

so

$$r = (\varepsilon^r_n \otimes 1_s) \circ (1_n \otimes \varepsilon^l_n \otimes 1_n \otimes 1_s \otimes \varepsilon^l_n).$$

Hence, the meaning of the sentence is given by:

$$\begin{aligned}&\mathcal{F}((\varepsilon^r_n \otimes 1_s) \circ (1_n \otimes \varepsilon^l_n \otimes 1_n \otimes 1_s \otimes \varepsilon^l_n))(\overrightarrow{carnivorous} \otimes \overrightarrow{animals} \otimes \overrightarrow{eat} \otimes \overrightarrow{meat})\\ &\quad = (\varepsilon_N \otimes 1_S) \circ (1_N \otimes \varepsilon_N \otimes 1_N \otimes 1_S \otimes \varepsilon_N)(\overrightarrow{carnivorous} \otimes \overrightarrow{animals} \otimes \overrightarrow{eat} \otimes \overrightarrow{meat}).\end{aligned}$$



From the diagram we can simply read off the meaning map, which is:

$$\varphi = (\varepsilon_N \otimes 1_S) \circ (1_N \otimes \varepsilon_N \otimes 1_N \otimes 1_S \otimes \varepsilon_N).$$

Suppose that the individual words in the sentence are given by:

$$\overline{carnivorous} = \sum_i C_i^{carni}\, \overrightarrow{n_i} \otimes \overrightarrow{n_i} \qquad \overrightarrow{animals} = \sum_k \alpha_k\, \overrightarrow{n_k}$$

$$\overrightarrow{meat} = \sum_j \beta_j\, \overrightarrow{n_j} \qquad \overline{eat} = \sum_{lrt} C^{eat}_{lrt}\, \overrightarrow{n_l} \otimes s_r \otimes \overrightarrow{n_t}$$

Then we obtain the meaning of the sentence as:

$$\begin{aligned}&\varphi\left(\overline{carnivorous} \otimes \overrightarrow{animals} \otimes \overline{eat} \otimes \overrightarrow{meat}\right).\\ &= \varphi\left(\sum_i C_i^{carni}\, \overrightarrow{n_i} \otimes \overrightarrow{n_i} \;\otimes\; \sum_k \alpha_k\, \overrightarrow{n_k} \;\otimes\; \sum_{lrt} C^{eat}_{lrt}\, \overrightarrow{n_l} \otimes s_r \otimes \overrightarrow{n_t} \;\otimes\; \sum_j \beta_j\, \overrightarrow{n_j}\right)\\ &= (\varepsilon_N \otimes 1_S)\left(\sum_{ijklrt} C_i^{carni}\alpha_k\, C^{eat}_{lrt}\beta_j\, \langle\overrightarrow{n_i}|\overrightarrow{n_k}\rangle\langle\overrightarrow{n_t}|\overrightarrow{n_j}\rangle\, \overrightarrow{n_i} \otimes \overrightarrow{n_l} \otimes s_r\right)\\ &= (\varepsilon_N \otimes 1_S)\left(\sum_{ijlr} C_i^{carni}\alpha_i\, C^{eat}_{lrj}\beta_j\, \overrightarrow{n_i} \otimes \overrightarrow{n_l} \otimes s_r\right)\\ &= \sum_{ijrl} C_i^{carni}\alpha_i\, C^{eat}_{lrj}\, \beta_j\, \langle\overrightarrow{n_i}|\overrightarrow{n_l}\rangle\, s_r\\ &= \sum_{ijr} C_i^{carni}\, \alpha_i\, C^{eat}_{irj}\, \beta_j\, s_r.\end{aligned}$$

2.4 Relative clauses via Frobenius algebras

We introduce very briefly the additional structure on top of compact closed categories developed in [34,35], which allows for modeling sentences and phrases containing relative pronouns *that*, *which*, *who*, *whose* and the possessive *whose*. This additional structure is provided by the so-called Frobenius algebras, first developed by F. G. Frobenius.

Definition 16 (Frobenius algebra [34])**.** *Let $\mathcal{C}$ be a symmetric monoidal category and $A \in Ob(\mathcal{C})$. A Frobenius algebra $\mathcal{A}$ over $\mathcal{C}$ is a tuple $(A,\, \Delta,\, \iota,\, \zeta,\, \mu)$ where:*

- $(A,\, \mu : A \otimes A \to A,\, \zeta : I \to A)$ *is an internal monoid;*
- $(A,\, \Delta : A \to A \otimes A,\, \iota : A \to I)$ *is an internal comonoid.*

These together satisfy the Frobenius condition:

$$(\mu \otimes 1_A) \circ (1_A \otimes \Delta) = \Delta \circ \mu = (1_A \otimes \mu) \circ (\Delta \otimes 1_A)\,.$$

For a definition of monoid and comonoid see, e.g. [3], and for more on Frobenius algebras refer to [20]. For our purposes we will only need to consider Frobenius algebras over **FHilb**.

Let $V \in Ob(\mathbf{FHilb})$ be a finite-dimensional (real) Hilbert space with basis $\{\overrightarrow{e_i}\}_i$. We define a Frobenius algebra over it by:

$$\begin{array}{llll}
\mu : V \otimes V \to V & \zeta : I \to V & \Delta : V \to V \otimes V & \iota : V \to I \\
\quad \overrightarrow{e_i} \otimes \overrightarrow{e_j} \mapsto \delta_{ij}\, \overrightarrow{e_i} & \quad 1 \mapsto \sum_i \overrightarrow{e_i} & \quad \overrightarrow{e_i} \mapsto \overrightarrow{e_i} \otimes \overrightarrow{e_i} & \quad \overrightarrow{e_i} \mapsto 1
\end{array}$$

The intuition behind using these to model language structures can be summarized according to [35] as:

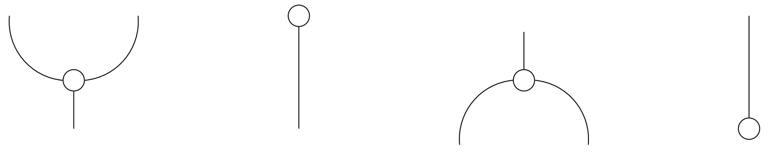
The comonoid's comultiplication Δ is often referred to as **copying**. It has the effect of producing a diagonal matrix out of a vector, i.e. $\Delta(\overrightarrow{v})$ is the diagonal matrix in $V \otimes V$ whose diagonal entries are the coefficients of $\overrightarrow{v}$. Copying enables the transfer of information contained in a single vector (or the vector space it belongs to) to two others. For example, by copying the information contained in a noun vector we can feed the two copies into a relative pronoun and a verb at the same time.

The monoid's multiplication μ is referred to as **uncopying**. It extracts the diagonal entries of a matrix and produces a vector with these as coefficients. Uncopying allows us to merge the information coming from two different sources. This can be used in cases where we need to put back together the information produced after processing different parts of a sentence containing a relative clause into a single output.

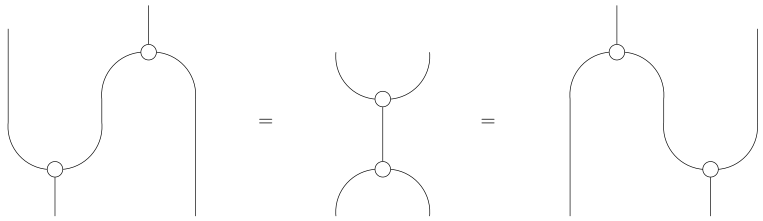
The unit ι lets us discard information. For example, a relative pronoun takes as input from the verb a wire of type S corresponding to sentence type s, but does not produce an output of the same type (as relative clauses output noun-types) and has to discard it.

Diagrams for Frobenius algebra morphisms.

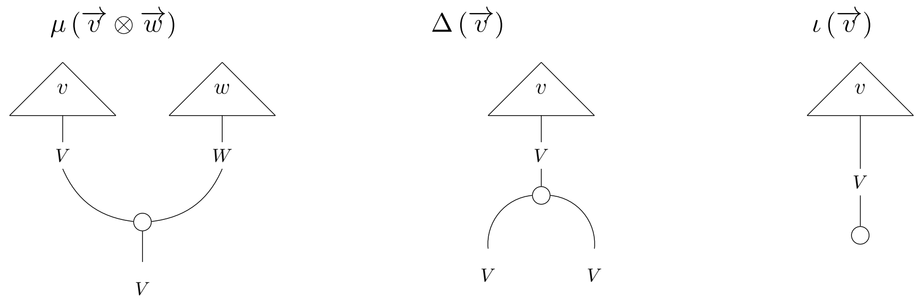
$\mu : A \otimes A \to A \qquad \zeta : I \to A \qquad \Delta : A \to A \otimes A \qquad \iota : A \to I$



The Frobenius condition is depicted as:



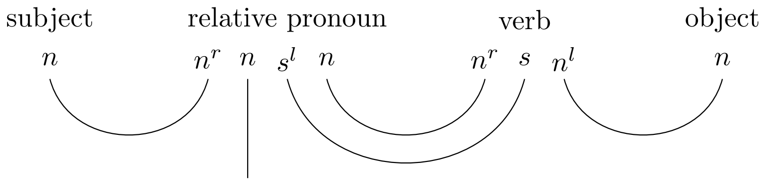
Applications of the Frobenius maps to vector are depicted as the composition of the Frobenius map with the corresponding state:


$\mu\,(\overrightarrow{v} \otimes \overrightarrow{w})$ $\quad\Delta\,(\overrightarrow{v})$ $\quad\iota\,(\overrightarrow{v})$

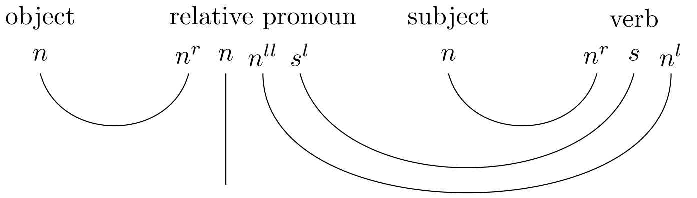
where $v : I \to V$ and $w : I \to W$.

2.4.1 Relative pronouns via Frobenius algebras

Recall that the grammatical type of the subject relative pronouns *who*, *that* and *which* is $n^r n s^l n$ and that of the object relative pronouns *whom*, *that* and *which* is $n^r n n^{ll} s^l$. The grammatical reductions of relative clauses are established as follows.



subject relative clause

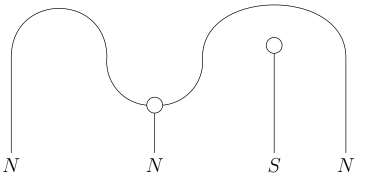


object relative clause

Note that the output in either case is of type n, as expected from a well-typed relative clause. Applying the functor $\mathcal{F} : \mathbf{Perg}_{\mathcal{B}} \to \mathbf{FHilb}$ to the types of the relative clauses gives us the tensor spaces in which their meanings live.

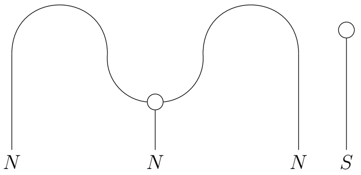
$$\mathcal{F}\left(n^r n s^l n\right) = N \otimes N \otimes S \otimes N$$
$$\mathcal{F}\left(n^r n n^{ll} s^l\right) = N \otimes N \otimes N \otimes S$$

According to [34], these are functional words explicitly given in **FHilb** as:



subject

$$(1_N \otimes \mu_N \otimes \zeta_S \otimes 1_N) \circ (\eta_N \otimes \eta_N)$$

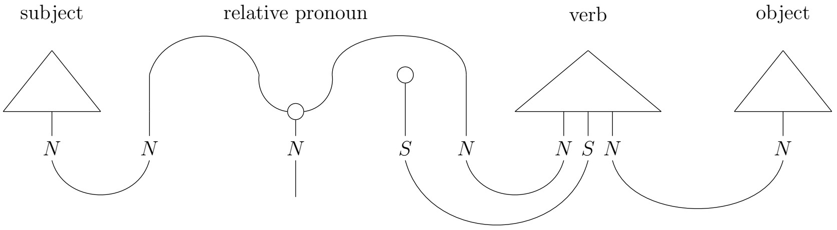


object

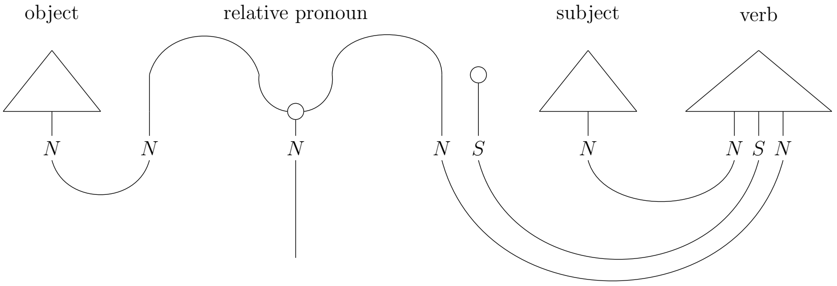
$$(1_N \otimes \mu_N \otimes 1_N \otimes \zeta_S) \circ (\eta_N \otimes \eta_N)$$

Then the diagrammatic representations for subject and object relative clauses become:

subject relative clause

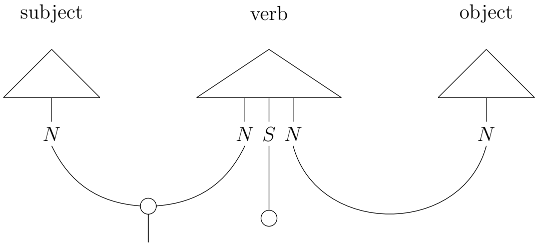


object relative clause



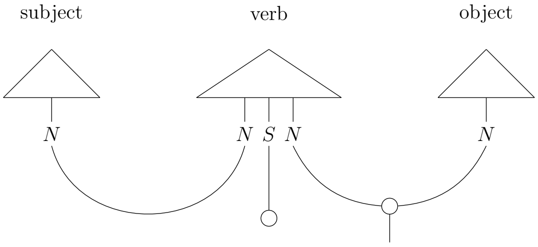
These reduce to:

subject relative clause



$$(\mu_N \otimes \iota_S \otimes \varepsilon_N) \left(\overrightarrow{subj} \otimes \overline{verb} \otimes \overrightarrow{obj}\right)$$

object relative clause



$$(\varepsilon_N \otimes \iota_S \otimes \mu_N) \left(\overrightarrow{subj} \otimes \overline{verb} \otimes \overrightarrow{obj}\right)$$

Note that these can now be used in more complicated structures, such as nested relative clauses or positive transitive sentences containing various subject or object relative clauses. For concrete examples, see [34]. We will use instances of this construction with in the framework of the CPM(**FHilb**) in the final chapter.

In their paper [35], Sadrzadeh, Clark and Coecke apply Frobenius algebras in a similar fashion in order to model meanings of possessive subject and object relative clauses, i.e. structures of the form **possessor whose subject verb object** and **possessor whose object subject verb**. These are not discussed here.

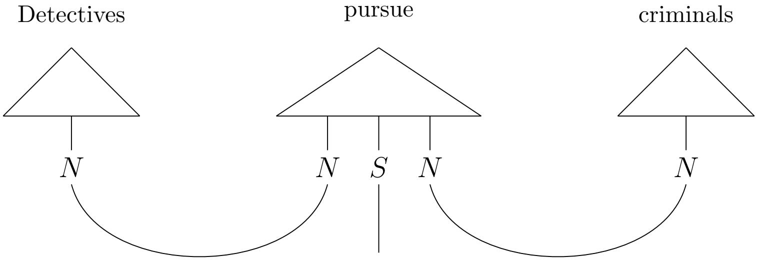
2.5 Concrete vector spaces for nouns and sentences

So far we haven't said much about how the vector spaces N and S are chosen and how the output of the meaning map can be interpreted so as to make the categorical compositional model of [7] truly distributional and practically applicable. A detailed discussion of the various options that have been utilized so far and experimental support is not necessary here, but the reader is referred to [12–14,36]. We will simply outline two main approaches that can be adopted and which are used later on in this thesis in computations of examples.

We will introduce a running example that will help illustrate the different approaches to choosing N and S and show the kind of results we obtain in practice. This will be a simple positive definite sentence, but a very similar approach can be adopted when representing adjective-noun phrases and more complicated structures involving various phrases and even adverbs [14]. The example will be **Detectives pursue criminals.** Without specifying the common noun space N and sentence space S, we define the individual word vectors as:

$$\overrightarrow{detectives} = \sum_i \alpha_i \overrightarrow{n_i} \qquad \overrightarrow{criminals} = \sum_j \beta_j \overrightarrow{n_j} \qquad \overrightarrow{pursue} = \sum_{prt} C_{prt}\overrightarrow{n_p} \otimes \overrightarrow{s_r} \otimes \overrightarrow{n_t}.$$

The meaning of the sentence is given by



$$\begin{aligned}
&(\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\left(\overrightarrow{detectives} \otimes \overrightarrow{pursue} \otimes \overrightarrow{criminals}\right)\\
&= (\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\left(\sum_i \alpha_i \overrightarrow{n_i} \otimes \sum_{prt} C_{prt}\overrightarrow{n_p} \otimes \overrightarrow{s_r} \otimes \overrightarrow{n_t} \otimes \sum_j \beta_j \overrightarrow{n_j}\right)\\
&= \sum_{iprtj} \alpha_i \, C_{prt} \, \beta_j \, \langle n_i \mid n_p \rangle \, s_r \, \langle n_t \mid n_j \rangle\\
&= \sum_{irj} \alpha_i \, C_{irj} \, \beta_j \, s_r
\end{aligned} \tag{2.1}$$

2.5.1 Truth-theoretic meaning

In [7], the compositional categorical model of meaning introduced in the paper is applied in a truth-theoretic setting, in which we assume that the noun vectors are all (standard) orthonormal basis vectors in a vector space N, chosen to be the span of a suitable set of vectors which can be as small as simply the set of those basis vectors that represent the nouns under consideration. The sentence space S is taken to be one- or two-dimensional and truth-theoretic. In the one-dimensional case, i.e. when $S = Span\{\overrightarrow{1}\}$ we can take the basis vector $\overrightarrow{1}$ to stand for *true* and $\overrightarrow{0}$ for *false.* Alternatively, we can work with $S = Span\{|0\rangle, |1\rangle\}$, in which case $|0\rangle = \begin{pmatrix}1\\0\end{pmatrix}$ stands for *true* and $|1\rangle = \begin{pmatrix}0\\1\end{pmatrix}$ for *false.*

In our example, we can take N to be the two dimensional vector space spanned by:

$$\overrightarrow{n_1} = \begin{pmatrix}1\\0\end{pmatrix} = \overrightarrow{detectives} \quad \text{and} \quad \overrightarrow{n_2} = \begin{pmatrix}0\\1\end{pmatrix} = \overrightarrow{criminals},$$

and S to be the one-dimensional vector space with $C_{irj}s_r = \begin{cases}\overrightarrow{1} & \text{if } \overrightarrow{n_i} \text{ pursues } \overrightarrow{n_j}\\ \overrightarrow{0} & \text{o.w.}\end{cases}.$

Then (2.1) returns $\overrightarrow{1}$ whenever it is true that detectives pursue criminals (according to our sources) and $\overrightarrow{0}$ otherwise. The same outcome can be achieved if we take S to be 2-dimensional. Also, note that if we increase the size of the noun space by, e.g., including basis vectors for all possible detectives and criminals and representing the vectors $\overrightarrow{detectives}$ and $\overrightarrow{criminals}$ as sums of these, the above calculation will result in a sum of $\overrightarrow{1}$'s and $\overrightarrow{0}$'s, where there will be as many $\overrightarrow{1}$'s in this sum as there are detective-criminal pairs of people in the relationship of the latter being pursued by the former. This can be generalised in other ways, such as by adding weights to the vectors that go into the sums corresponding to the nouns, or even by representing the verb as a sum of other actions, e.g.:

$$\overrightarrow{pursue} = \frac{1}{2}\overrightarrow{investigate} + \frac{1}{2}\overrightarrow{chase}.$$

For worked examples of this form see [7].
This framework is useful for a limited number of applications and for proof of concept examples, but is otherwise essentially no different from the model-theoretic view on semantics. For a truly *distributional* perspective, we need to extract the word meanings from a corpus. Moreover, in the case where the output of the meaning map is of type s, taking S to be truth-theoretic does not tell us anything about the features of the individual words that comprise the sentence and how these interact with each other to produce the output. As such, it is not a very useful space to work with when considering practical tasks, or even for sentence comparisons.

2.5.2 A more distributional approach to word meaning and a new sentence space

The idea developed in [14] is to take N to be a structured vector space like in other applications of the distributional models framework. The basis vectors of this space can be taken to be a set of properties or salient features of the nouns, depending on the application we are interested in. In the case of applications in the area of cognitive linguistics, such as [25], these basis vectors can be taken to be the salient features of the nouns that do not necessarily have to be extracted from a corpus, but can be obtained from target groups in controlled experiments. We will see an example of this in Chapter 4.

In cases where we are interested in practical linguistic applications (definitions, ambiguity, information retrieval, sentence comparisons, etc.) and working with a corpus, these vectors can be obtained directly from the corpus via statistical (co-occurrence) methods. In this thesis, we will use the approach of [14] for the purposes of toy examples whenever this is more suitable than a simple truth-theoretic model. We will illustrate how this approach works by means of an example. Suppose the following are basis vectors for N: *arg-strong*, *subj-build*, *obj-clean.* Then if we want to represent a noun $\overrightarrow{n}$ with respect to these basis vectors, it will have entires that express how many times in the corpus this noun has appeared as the argument of the adjective *strong*, the subject of the verb *build* and the object of the verb *clean.* The sentence space can be taken to be $S = N \otimes N$, so that the basis vectors for S are of the form $\overrightarrow{n_i} \otimes \overrightarrow{n_j}$, where $\overrightarrow{n_i}, \overrightarrow{n_j}$ are basis elements of N. Verbs are then represented as:

$$\overrightarrow{verb} = \sum_{ij} C_{ij}^{verb}\, \overrightarrow{n_i} \otimes (\overrightarrow{n_i} \otimes \overrightarrow{n_j}) \otimes \overrightarrow{n_j}.$$

The intuition behind this is that it allows us to output a meaning of the sentence in the form of a matrix that contains all the information that we get from modifying the features of the noun object and subject via the verb. The weights C_{ij}^{verb} are built by counting the number of times that a word which is a possessor of property n_i is the subject of this verb and a word that has property n_j is its object, i.e. how many times we have in the corpus *(something which is n_i) verb (something which is n_j).*

For example, suppose that we have (standard) basis vectors $\{n_1, \dots n_5\}$ representing, in this order, *arg-cunning*, *arg-righteous*, *arg-nasty*, *arg-meticulous*, *obj-kill* and in our imaginary corpus the word *detective* has appeared 4 times as the argument of *cunning*, 5 times as the argument of *righteous*, 3 times as the argument of *meticulous* and once as the object of *kill.* Suppose that the word *criminal* has appeared twice as the argument of *cunning* and *meticulous*, 3 times as the argument of *nasty.*

Then the vectors for the two words are given by:

$$\overrightarrow{detective} = \begin{pmatrix} 4 \\ 5 \\ 0 \\ 3 \\ 1 \end{pmatrix}, \quad \overrightarrow{criminal} = \begin{pmatrix} 2 \\ 0 \\ 3 \\ 2 \\ 0 \end{pmatrix},$$

and the meaning of the sentences can be computed accordingly.

Chapter 3

The CPM construction and density matrix formalism

The model of meaning of [7] summarised in the previous chapter has been applied successfully to a number of tasks [11–14, 25, 36]. However, restricting our attention to vectors as means of capturing word meaning also has limitations, which prevent us from making good use of it for some more involved applications that necessitate for the relationships between words and between concepts and their salient features to be taken into account. For this reason, some of the more recent work in the field [2, 32, 33] has involved a shift of focus from vectors to density matrices to represent semantics. This approach has already proved to be promising for disambiguation [33] and in the modeling to hyponymy [2] and has given rise to many opportunities for further research.

Density matrices are used in quantum mechanics to represent uncertainty about the state of a physical system. Not only do they possess richer structure than vectors, but they are also equipped with mathematical properties that allow for the introduction of various symmetric and asymmetric measures of similarity and inclusion. In contrast, vectors can only be compared to each other via symmetric measures, such as the cosine.

The mathematical framework in which density matrices occupy the same position as that of vectors with respect to the †-compact closed category **FHilb** is the category CPM(**FHilb**), built on top of **FHilb**. Thus, in this chapter we will first introduce the CPM construction in the context of its applications to distributional models of meaning, following closely [33], and then shift our attention to CPM(**FHilb**) and density matrices specifically.

3.1 The doubling construction and CPM

3.1.1 The doubling construction on †-compact closed categories

Definition 17 (The (doubling) $\mathbf{D}$ construction)**.** *The* ***D*** *construction* $\mathbf{D}(\mathcal{C})$ *on a symmetric* †*-compact closed category* $\mathcal{C}$ *with monoidal tensor* $\otimes$ *is given as follows:*

- Objects: *The objects* $Ob(\mathbf{D}\mathcal{C})$ *of* $\mathbf{D}(\mathcal{C})$ *are the same as the objects* $Ob(\mathcal{C})$ *or* $\mathcal{C}$*.*
- Morphisms: *The morphisms* $Ar(\mathbf{D}\mathcal{C})$ *of* $\mathbf{D}(\mathcal{C})$ *consist of* $\varphi \in \mathbf{D}\mathcal{C}(A, B)$*, where*

  $$\varphi \in \mathcal{C}(A^* \otimes A, B^* \otimes B).$$

  In other words, $\varphi : A \to B$ *is a morphism in the double category if* $\varphi : A^* \otimes A \to B^* \otimes B$ *is a morphism in* $\mathcal{C}$*.*
- Morphism composition: *If* φ*,* $\psi \in Ar(\mathbf{D}\mathcal{C})$ *are two morphisms of* $\mathbf{D}(\mathcal{C})$ *with* $\varphi \in \mathbf{D}\mathcal{C}(A, B)$ *and* $\psi \in \mathbf{D}\mathcal{C}(B, C)$*, then the composition* $\psi \circ \varphi \in \mathbf{D}\mathcal{C}(A, C)$ *is provided by the morphism:*

  $$\psi \circ \varphi : A^* \otimes A \to C^* \otimes C\,,$$

whose existence follows from the composition of $\varphi : A^* \otimes A \to B^* \otimes B$ *and* $\psi : B^* \otimes B \to C^* \otimes C$ *in* $\mathcal{C}$*.*

- Monoidal tensor: *The monoidal tensor of* $\mathbf{D}(\mathcal{C})$ *is given by* $\otimes_D : \mathbf{D}(\mathcal{C}) \to \mathbf{D}(\mathcal{C})$*, which acts*
  - on objects $A, B \in Ob(\mathbf{D}\mathcal{C})$ *as* $A \otimes_D B = A \otimes B$ *;*
  - on morphisms $\varphi \in \mathbf{D}\mathcal{C}(A, B)$, $\psi \in \mathbf{D}\mathcal{C}(C, D)$ *:*

$$\varphi \otimes_D \psi : A^* \otimes A \otimes C^* \otimes C \xrightarrow{\varphi \otimes \psi} B^* \otimes B \otimes D^* \otimes D$$

Note that the existence of the swap map $\sigma : A \otimes B \xrightarrow{\cong} B \otimes A$ *in* $\mathcal{C}$ *implies that:*

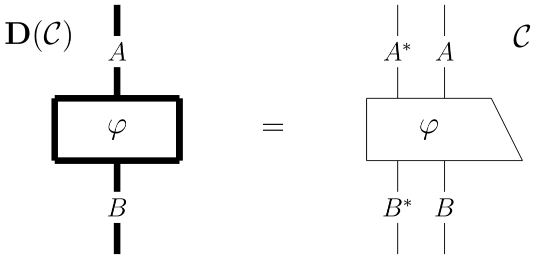
$$A^* \otimes A \otimes C^* \otimes C \cong A^* \otimes C^* \otimes C \otimes A \quad \textit{and} \quad B^* \otimes B \otimes D^* \otimes D \cong B^* \otimes D^* \otimes D \otimes B,$$

so we take:

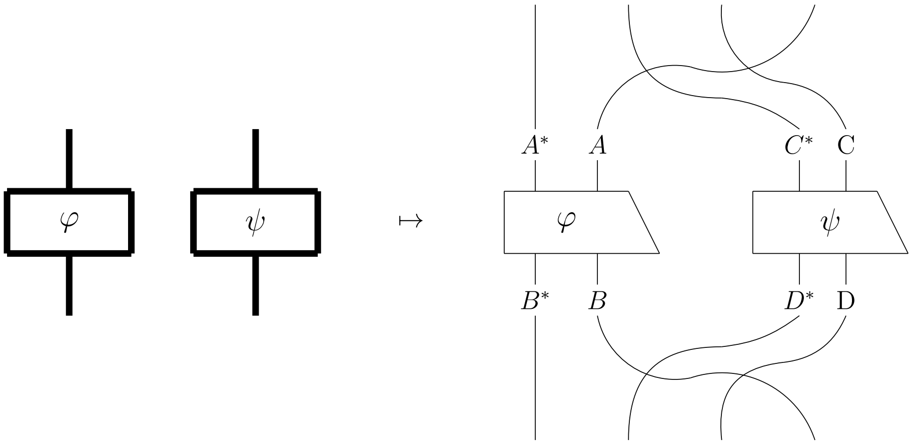
$$\varphi \otimes_D \psi : A^* \otimes C^* \otimes C \otimes A \xrightarrow{\varphi \otimes \psi} B^* \otimes D^* \otimes D \otimes B.$$

Graphical calculus in $\mathbf{D}\mathcal{C}$

Our convention will be to represent the boxes and wires of $\mathbf{D}(\mathcal{C})$ with thicker lines so as to distinguish between the structures $\mathbf{D}(\mathcal{C})$ and $\mathcal{C}$. This is done because the transfer from the latter category to the former essentially means doubling of the wires. Thus, for example, the map $\varphi \in \mathbf{D}\mathcal{C}(A, B)$ is depicted as the LHS part of the diagram below, while its corresponding counterpart in $\mathcal{C}$, $\varphi : A^* \otimes A \to B^* \otimes B$, as the RHS:



The tensor of two morphisms φ and ψ is depicted as:



$\mathbf{D}(\mathcal{C})$ is also a †-compact closed category [37]. It inherits this structure from $\mathcal{C}$ via a strict monoidal functor $E : \mathcal{C} \to \mathbf{D}(\mathcal{C})$ given by:

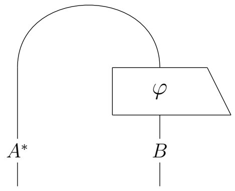
$$\begin{aligned} &E : Ob(\mathcal{C}) \to Ob(\mathbf{D}\mathcal{C}) && E : Ar(\mathcal{C}) \to Ar(\mathbf{D}\mathcal{C}) \\ &E :: A \mapsto A && E :: f \mapsto f_* \otimes f \end{aligned}$$

and, inductively, $E(\varphi \otimes \psi) = E(\varphi) \otimes_D E(\psi)$.

There is a bijective correspondence between the states of $\mathbf{D}(\mathcal{C})$ i.e. morphisms $\varphi : I \to A^* \otimes A$ of $\mathcal{C}$ and the positive operators on A, i.e. morphisms $\psi : A \to A$ of $\mathcal{C}$, established via the map that sends each operator to its name:

$$\begin{aligned} f : &\mathcal{C}(A, A) \to \mathcal{C}(I, A^* \otimes A) \\ &\varphi \mapsto \ulcorner \varphi \urcorner = (id_{A^*} \otimes \varphi) \circ \eta_A \end{aligned}$$

These states will be of central importance once we start working in the category CPM(**FHilb**). Recall that the name of the operator $\varphi : I \to A^* \otimes B$, $\ulcorner \varphi \urcorner$ is depicted (in $\mathcal{C}$) as follows:

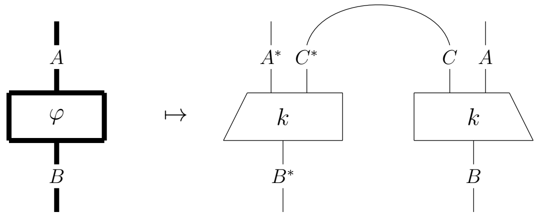


3.1.2 The subcategory CPM($\mathcal{C}$) of $\mathbf{D}(\mathcal{C})$

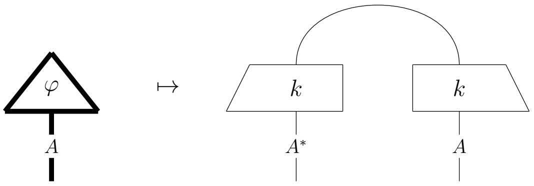
Before we define the CMP($\mathcal{C}$) subcategory of $\mathbf{D}(\mathcal{C})$ we need the notion of a completely positive map. Completely positive maps arise naturally in the context of density matrices as they are essentially density matrix-preserving maps. This concept will be made more precise later, but for now we give a general definition of completely positive maps that works in any †-compact closed category. The definition below is due to Selinger [37].

Definition 18 (Completely positive morphism [33]). *Let $\varphi \in \mathbf{D}\mathcal{C}(A, B)$ be a morphism of $\mathbf{D}(\mathcal{C})$, i.e. $\varphi : A^* \otimes A \to B^* \otimes B$. We say that φ is a* completely positive *morphism if there exists $C \in Ob(\mathbf{D}\mathcal{C})$ and $k \in \mathcal{C}(C \otimes A, B)$, such that φ can be embedded in $\mathcal{C}$ by:*

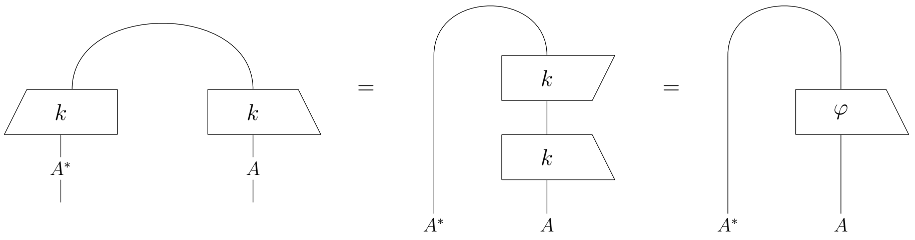
$$\varphi \mapsto (k_* \otimes k) \circ (1_{A^*} \otimes \eta_C \otimes 1_A)\,.$$



Note that this implies that states in CPM($\mathcal{C}$) can be represented as:



and we can recover the original definition from above via:



The second equality follows from the standard definition of *positive operator*, which tells us that φ being positive means that we can express it as $\varphi = k \circ k^{\dagger}$. The LHS diagram will be the one used in applications.

This representation implies that if f and g are two completely positive morphisms then the following are also completely positive:

$$f \circ g\,, \qquad f \otimes g\,, \qquad f_* \otimes f = E(f). \tag{3.1}$$

This essentially tells that any morphism in the category $\mathbf{D}(\mathcal{C})$ that can be obtained via a combination of tensors and compositions of other completely positive morphisms is also completely positive. Thus, we have closure under $\circ$ and $\otimes$ of completely positive morphisms in $\mathbf{D}(\mathcal{C})$, and hence can define the following subcategory.

Definition 19 (CPM($\mathcal{C}$))**.** *If $\mathcal{C}$ is a $\dagger$-compact closed category then CPM($\mathcal{C}$) is the subcategory of $\mathbf{D}(\mathcal{C})$ which has the same objects as $\mathcal{C}$ and its morphisms are the completely positive morphisms of $\mathbf{D}(\mathcal{C})$.*

Let I be the embedding of CPM($\mathcal{C}$) into $\mathbf{D}(\mathcal{C})$. Then there exists (by (3.1)) a strictly monoidal functor $\tilde{E} : \mathcal{C} \to$ CPM($\mathcal{C}$) such that $E = I\tilde{E}$.

The compact closure and Frobenius maps in CPM($\mathcal{C}$)

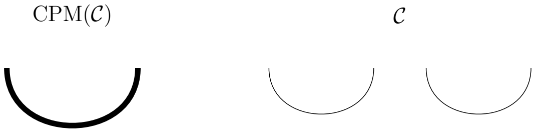
In the category CPM($\mathcal{C}$) the compact closure maps ε and η are given by $E(\varepsilon)$ and $E(\eta)$, and similarly the Frobenius algebra maps ι, ζ, μ and Δ are given my $E(\iota)$, $E(\zeta)$, $E(\mu)$ and $E(\Delta)$. Note that whenever there is no ambiguity as to which category we are working in, we will adopt the convention of writing ε, η, μ, Δ, ι and ζ to mean the corresponding morphism in $\mathcal{C}$ or in CPM($\mathcal{C}$). Also, in this context we are normally only interested in left adjoints, so we will write for each $A \in Ob(\mathbf{D}\mathcal{C})$:

$$\varepsilon^l_A = \varepsilon : A^* \otimes A \to I \qquad\qquad \eta^l_A = \eta^* : I \to A \otimes A^*$$

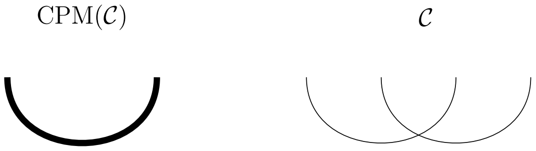
Later on, when working with real Hilbert spaces, it will not matter which map we mean, as the left and right vector space adjoints are both isomorphic to the space itself.

Diagrammatic Calculus for CPM($\mathcal{C}$)

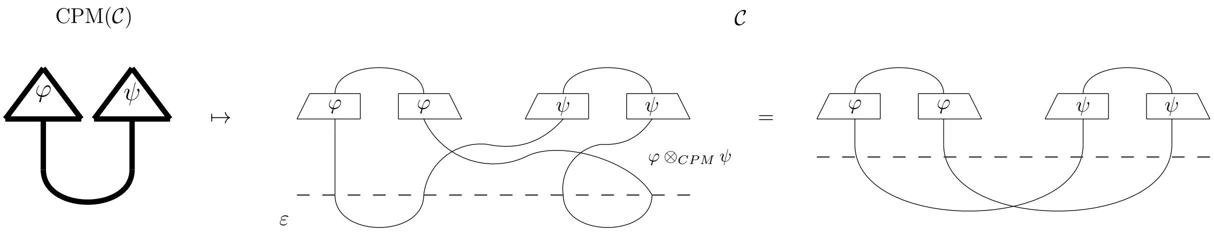
We will mainly be interested here in how the structure-preserving maps combine with the states of the category, so a brief discussion about this is in place. First of all, note that the diagrams for all of the structural morphisms and Frobenius maps are simply obtained from their original counterparts by doubling the wires. For example, for the ε-map in the CPM category, we get:



However, applying the map to the tensor of two states in CPM results in some swaps in the output wires that occur because of the way that the tensor of morphisms is defined in CPM (see diagram above). It will be more convenient for us to represent this in an alternative, but equivalent fashion, whereby we can treat, for example, the ε-morphism as being:



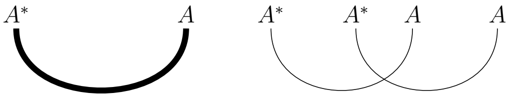
To see how this works, consider the following diagram, in which on the LHS we have the diagrammatic representation of $\varepsilon\,(\varphi \otimes \psi)$ in CPM($\mathcal{C}$), in the middle we have the corresponding diagram in $\mathcal{C}$ and on the RHS - the equivalent representation that will be used in applications.



Thus, we can summarise these morphisms in CPM($\mathcal{C}$) via the following diagrams:

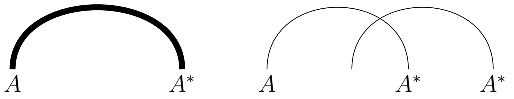
CPM($\mathcal{C}$) $\mathcal{C}$ CPM($\mathcal{C}$) $\mathcal{C}$

$E(\varepsilon) = \varepsilon_* \otimes \varepsilon$ $\quad \varepsilon : A^* \otimes A^* \otimes A \otimes A \to I$



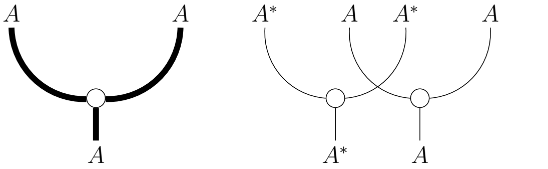
$$\varepsilon : (\overrightarrow{e_i} \otimes \overrightarrow{e_j}) \otimes (\overrightarrow{e_k} \otimes \overrightarrow{e_l}) \mapsto \langle \overrightarrow{e_i} \mid \overrightarrow{e_k} \rangle \langle \overrightarrow{e_j} \mid \overrightarrow{e_l} \rangle$$

$E(\eta) = \eta_* \otimes \eta$ $\quad \eta : I \to A \otimes A \otimes A^* \otimes A^*$



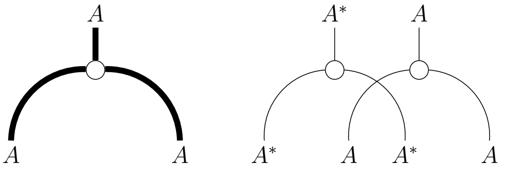
$$\eta : 1 \mapsto \sum_{ij} \overrightarrow{e_i} \otimes \overrightarrow{e_j} \otimes \overrightarrow{e_i} \otimes \overrightarrow{e_j}$$

$E(\mu) = \mu_* \otimes \mu$ $\quad \mu : A^* \otimes A \otimes A^* \otimes A \to A^* \otimes A$



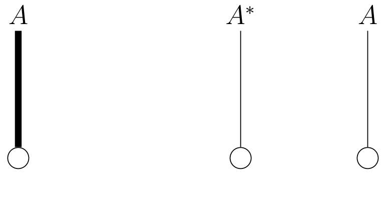
$$\mu : (\overrightarrow{e_i} \otimes \overrightarrow{e_j}) \otimes (\overrightarrow{e_k} \otimes \overrightarrow{e_l}) \mapsto \langle \overrightarrow{e_i} \mid \overrightarrow{e_k} \rangle \langle \overrightarrow{e_j} \mid \overrightarrow{e_l} \rangle (\overrightarrow{e_i} \otimes \overrightarrow{e_j})$$

$E(\Delta) = \Delta_* \otimes \Delta$ $\quad \Delta : A^* \otimes A \to A^* \otimes A \otimes A^* \otimes A$



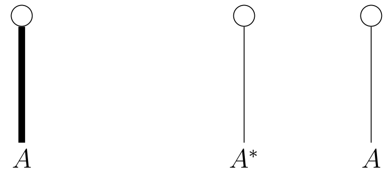
$$\Delta : (\overrightarrow{e_i} \otimes \overrightarrow{e_j}) \mapsto \overrightarrow{e_i} \otimes \overrightarrow{e_j} \otimes \overrightarrow{e_i} \otimes \overrightarrow{e_j}$$

$E(\iota) = \iota_* \otimes \iota$ $\quad \iota : I \to A^* \otimes A$



$$\iota : (\overrightarrow{e_i} \otimes \overrightarrow{e_j}) \mapsto 1$$

$E(\zeta) = \zeta_* \otimes \zeta$ $\quad \zeta : A^* \otimes A \to I$



$$\zeta : 1 \mapsto \sum_{i} \overrightarrow{e_1} \otimes \overrightarrow{e_i}$$

3.1.3 Sentence meaning in the category CPM($\mathcal{C}$)

The CPM construction allows us to consider a number of new candidate categories in place of **FHilb** for storing word meanings. As already mentioned, this thesis will only make use of CPM(**FHilb**), but since it is possible to work in other categories, such as CPM(*Rel*), we first define our sentence meaning map in a more general setting where we do not explicitly specify the category that we wish to use, but rather work with a general †-compact closed category $\mathcal{C}$ and assume that our words exist as states in the category CPM($\mathcal{C}$).

We now have all the necessary ingredients to define a new from-the-meaning-of-words-to-the-meaning-of-sentences map, or meaning map, that will allow us to transfer the grammatical structure of a sentence to a category containing its words' semantics, which is some CPM($\mathcal{C}$). Recall that in the vector space model of distributional models of meaning the transition between syntax and semantics was achieved via a strongly monoidal functor $\mathcal{F} : \mathbf{Preg}_{\mathcal{B}} \to \mathbf{FHilb}$. It turns out that we can similarly define a strongly monoidal functor $\mathcal{S} : \mathbf{Preg}_{\mathcal{B}} \to \mathrm{CPM}(\mathcal{C})$. We define this functor to be $S = \tilde{E}Q$, where Q is the strongly monoidal functor $Q : \mathbf{Preg}_{\mathcal{B}} \to \mathcal{C}$, of which $\mathcal{F} : \mathbf{Preg}_{\mathcal{B}} \to \mathbf{FHilb}$ is a special case. The morphisms involved can be summarised via the following diagram:

$$\begin{tikzcd}
 & \mathrm{CPM}(\mathcal{C}) \arrow[r, hook, "I"] & \mathbf{D}(\mathcal{C}) \\
\mathbf{Preg}_{\mathcal{B}} \arrow[ur, "\tilde{E}Q"] \arrow[r, "Q"] & \mathcal{C} \arrow[u, "\tilde{E}"] \arrow[ur, "E"'] &
\end{tikzcd}$$

Since Q is a strongly monoidal functor and $\tilde{E}$ is strictly monoidal, we get that S is strongly monoidal, as required. We can now define the meaning map that makes use of $\mathrm{CPM}(\mathcal{C})$ as follows.

Fix a †-compact closed category $\mathcal{C}$ and its corresponding CPM construction $\mathrm{CPM}(\mathcal{C})$ with unit object I. With the same notation as above we have:

Definition 20. *Let $s = w_1 \dots w_n$ be a string of words and let t_i be the grammatical type of word w_i in $\mathbf{Preg}_{\mathcal{B}}$. Suppose that the type reduction of s is given by $t_1 \dots t_n \xrightarrow{r} x$ for some $x \in Ob(\mathbf{Preg}_{\mathcal{B}})$. Let $\rho(w_i)$ be the meaning of word w_i in CPM($\mathcal{C}$), i.e. a state of the form $I \to S(t_i)$. Then the meaning of s is given by:*

$$S(r)\left(\rho(w_1) \otimes_{CPM} \dots \otimes_{CPM} \rho(w_n)\right). \tag{3.2}$$

3.2 Modeling word and sentence meaning in CPM(FHilb)

From now on, we will only be working with the category CPM(**FHilb**) that is built out of the already familiar category of finite dimensional Hilbert spaces and linear maps.

3.2.1 Density matrices as states in CPM(FHilb)

Recall that a *state* in a category $\mathcal{C}$ is a morphism of the form $\psi : I \to A$ for a vector space A, and that the states in **FHilb** are in a one-to-one correspondence with the elements of the vector space in question.

Definition 21. *A* pure state *on a vector space V is an operator $V \to V$ which is of the form $\varphi \circ \varphi^\dagger$, where $\varphi : I \to V$ is a state and $\varphi^\dagger \circ \varphi = id_I$.*

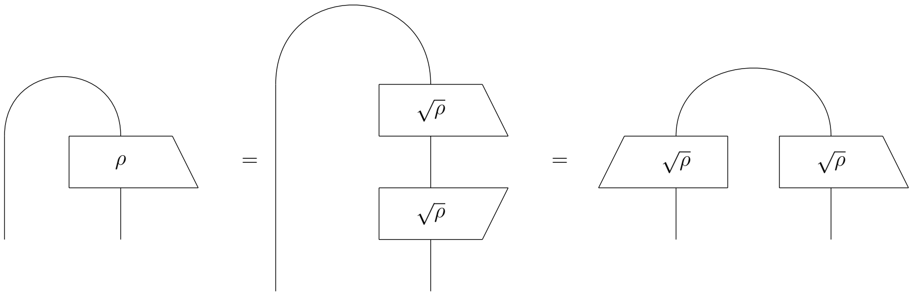
In quantum computing pure states represent the possible states in which a physical system can be. However, it is often the case that an observer does not have information about the exact state in which the system is, but rather only knows the probabilities attached to several possible states. This can be mathematically expressed as a convex sum of pure states and we will call this a mixed state. More precisely, we use density matrices.

Suppose that a system can exist in state $|\rho_i\rangle$ with probability p_i, for some collection of state-probability pairs $\{|\rho_i\rangle, p_i\}$. Then the *density matrix* or *density operator* corresponding to this system is given by:

$$\rho = \sum_i p_i \, |\rho_i\rangle\langle\rho_i|,$$

where $\sum_i p_i = 1$. Each of the $|\rho_i\rangle\langle\rho_i|$ is a pure state.
To see how these operators fit into our categorical framework, consider the following definition.

Definition 22 (Positive matrix [18])**.** *We call a matrix a* positive matrix *if it is the name of a positive morphism $\rho : V \to V$, i.e. a morphism $\ulcorner \rho \urcorner : I \to V^* \otimes V$. The morphism ρ is called a* mixed state. *Recall that these can be expressed diagrammatically as:*



Thus, a density matrix, which is a positive matrix with trace 1, is exactly a state in CPM(**FHilb**) – the category which has as objects finite-dimensional Hilbert spaces and as morphisms completely positive maps. Note that completely positive maps in this context means morphisms that take density matrices to density matrices and preserve their structure. All the formalism from the previous section carries over to the category CPM(**FHilb**). Just as before, we will only be using real-valued vector spaces and hence we will be able to assume that for all meaning spaces involved we have $V^* \cong V$.

3.2.2 Using density matrices to model word meanings

How can density matrices be used to capture word meaning and what do we gain by doing this, as opposed to sticking to vector-based representations?

One use of density matrices that mirrors their role in quantum computing is in representing words as probabilistic mixtures of their possible ambiguous meanings. For example, in [33], the ambiguous noun *queen* has the density matrix representation:

$$\ulcorner queen \urcorner = |Elisabeth\rangle\langle Elisabeth| + |band\rangle\langle band| + |chess\rangle\langle chess|,$$

where *Elisabeth*, *chess* and *band* are assumed to be all the possible meanings of the word *queen* and, furthermore, these are assumed to be themselves pure states, i.e. they have unambiguous meanings. The idea behind using this kind of representation is that once the ambiguous word is put in a sentence, the functional words in this sentence interact with it in a similar way to how an observation affects a physical system. This allows for a single meaning to emerge out of the collection of possible meanings and for this meaning to connect with the rest of the sentence and produce the relevant output, i.e. sentence meaning.

Another possible use of density matrices is in representing collective nouns as sums of their parts. For example, we could have that:

$$\ulcorner pet \urcorner = \sum_i p_i \ulcorner pet_i \urcorner,$$

where $\ulcorner pet_i \urcorner = |pet_i\rangle\langle pet_i|$ is the pure state corresponding to the i^{th} pet (e.g. $\ulcorner cat \urcorner$, $\ulcorner dog \urcorner$, etc). The advantage of doing this is that it lets us compare the collective nouns with their parts and see connections and differences between them that are not immediately obvious when using vectors. Also, it allows for the introduction of various asymmetric measures which facilitate the comparison and ordering of concepts and their components. Note that the same idea can be used for representing verbs or, indeed, any functional word in CPM(**FHilb**).

To see how these density matrices are formed in practice and how the morphisms work in the CPM(**FHilb**) category to form meaning maps and produce the meanings of sentences, consider the following example, where for simplicity all the nouns are assumed to be pure states.

Example

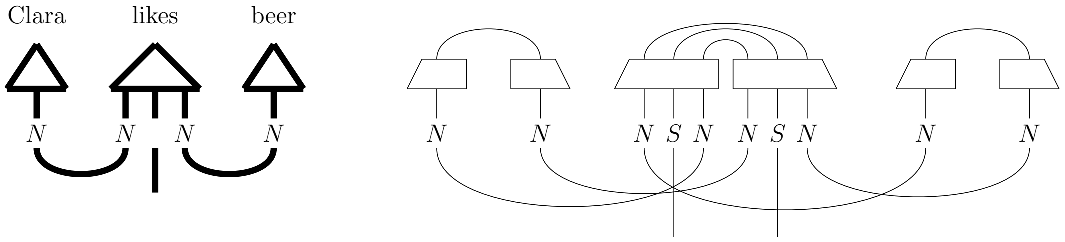
Let the noun space be given by a real Hilbert space N with basis vectors given by $\{|n_i\rangle\}_i$, where for some i, $|n_i\rangle = \overrightarrow{Clara}$ and for some j, $|n_j\rangle = \overrightarrow{beer}$. Let the sentence space be some unspecified S with basis $\{|s_i\rangle\}_i$. Then the density matrices for the nouns *Clara* and *beer* are given by:

$$\ulcorner Clara \urcorner = |n_i\rangle\langle n_i| \qquad \text{and} \qquad \ulcorner beer \urcorner = |n_j\rangle\langle n_j|.$$

Suppose the verb $\overrightarrow{like} \in N \otimes S \otimes N$ is given by $\overrightarrow{like} = \sum_{rtv} C_{rtv}\,|n_r\rangle|n_t\rangle|n_v\rangle$. Then its corresponding density matrix in CPM(**FHilb**) is given by:

$$
\begin{aligned}
\ulcorner like \urcorner &= \left(\sum_{rtv} C_{rtv}\,|n_r\rangle|n_t\rangle|n_v\rangle\right)\left(\sum_{klu} C_{klu}|n_k\rangle|n_l\rangle|n_u\rangle\right)^{T} \\
&= \sum_{rtvklu} C_{rtv}\,C_{klu}\,|n_r\rangle\langle n_k| \otimes |n_t\rangle\langle n_l| \otimes |n_v\rangle\langle n_u|
\end{aligned}
$$

The meaning map is simply $(\varepsilon_N \otimes 1_S \otimes \varepsilon_N)$ applied to $(\ulcorner Clara \urcorner \otimes \ulcorner like \urcorner \otimes \ulcorner beer \urcorner)$, as per the diagram below:



$$
\begin{aligned}
\rho(\text{Clara likes beer}) &= (\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\,(\ulcorner Clara \urcorner \otimes \ulcorner like \urcorner \otimes \ulcorner beer \urcorner) \\
&= (\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\left(|n_i\rangle\langle n_i| \;\otimes \sum_{klurtv} C_{rtv}C_{klu}\,|n_r\rangle\langle n_k| \otimes |s_t\rangle\langle s_l| \otimes |n_v\rangle\langle n_u| \;\otimes\; |n_j\rangle\langle n_j|\right) \\
&= \sum_{klurtv} C_{rtv}C_{klu}\,\langle n_i \,|\, n_r\rangle\,\langle n_i \,|\, n_k\rangle\,(|s_t\rangle\langle s_l|)\,\langle n_v \,|\, n_j\rangle\,\langle n_u \,|\, n_j\rangle \\
&= \sum_{ijtl} C_{itj}\,C_{ilj}\,|s_t\rangle\langle s_l|
\end{aligned}
$$

Chapter 4

Applications of distributional compositional models to cognitive linguistics phenomena

In this chapter we consider applications of the DisCoCat model of meaning to two cognitive linguistics phenomena, both of which possess some sort of asymmetry.

The so-called Pet Fish phenomenon is a classic example of overextension with respect to concept combination. We first give an overview of the recent work by Lewis and Coecke [25], who modeled this example in the original setting, in which words are represented by vectors. We then go on to consider the same in the new framework of CPM(**FHilb**).

The phenomenon of asymmetry in similarity judgments is examined in the classic experiments by A. Tversky [40], in which the perceived similarity of a more prominent country to a similar, but less prominent one is shown to be greater than the reverse. We present a simple method to capture this and mention an alternative solution that will be revisited in the next chapter.

4.1 Concept combination and the Pet Fish phenomenon

4.1.1 What is concept combination?

Concept combination relates to the way in which the meaning of the constituent parts of a phrase are connected to the meaning of the whole. For example, consider the simple adjective-noun phrase *fluffy cat* and suppose that we are interested in the connection between *fluffy*, *cat* and *fluffy cat.* Intuitively, this should be relatively straightforward - a fluffy cat is a concept that lies in the intersection of fluffy things and things which are cats. This is easily modeled with conjunction in classical set theory. However, if we tried to apply the same logic to the combination of concepts *school* and *furniture* into *school furniture*, then this approach does not yield intuitive results. *School furniture* are not things that are both *school* and *furniture*, but *furniture* which are being modified by the concept of school in some fashion.

Here we will be interested in the problem of typicality rating and membership judgment with respect to concept combination. Note that the membership problem can be treated simply as a special case of typicality. Informally speaking, this problem can be phrased as follows:

Given two concepts A and B and their combination AB, what can we conclude about the typicality (or membership) of an item x in A and in B from its typicality (membership) in AB, and vice versa?

These questions have been explored in detail since the 1980's both from a mathematical point of view and in experiments in psychology and some interesting and somewhat counterintuitive results have been observed. For example, in [17], J. Hampton concluded that human subjects from a target group considered some items to be members of the combined concept *school furniture* but not members of either the *school* or *furniture* categories. Similar results have been observed in connection with typicality ratings of items with respect to concepts and conjunctions of these concepts.

4.1.2 Overextension, the Pet Fish phenomenon and the shortfalls of fuzzy set theory

Here we will only look at the phenomenon of *overextension*, in which an item is perceived to have a higher degree of membership/typicality with respect to the combination of two concepts than to each of the concepts individually. A classical example of the overextension phenomenon in typicality judgment with respect to concept combination was cited way back in 1981 in [29] and is known as the Pet Fish phenomenon, or the guppy effect. Here a goldfish (or guppy) is judged to be a more typical representative of the concept *pet fish* than it is of either the class of pets or that of fish.

In fact, the authors argue that the typicality of an item with respect to a combination of two concepts cannot be determined via a simple logic function and that, in particular, treating the combination of concepts as their conjunction does not lead to fruitful results. That is to say, fuzzy set theory, which has been traditionally used to model concept combination, cannot efficiently be combined with prototype theory, since it leads to paradoxes in which an item is more prototypical of a conjunction of two concepts than of either of them.

In fuzzy set theory, the typicality rating of an item x with respect to a concept C is given by a membership function $f_C(x)$ and the typicality of x with respect to the combination of concepts C_1 and C_2 is given by its typicality w.r.t. their conjunction $C_1 \wedge C_2$, i.e. by $f_{C_1 \wedge C_2}(x)$, which satisfies the rule:

$$f_{C_1 \wedge C_2}(x) \leq min\{f_{C_1}(x), f_{C_2}(x)\}.$$

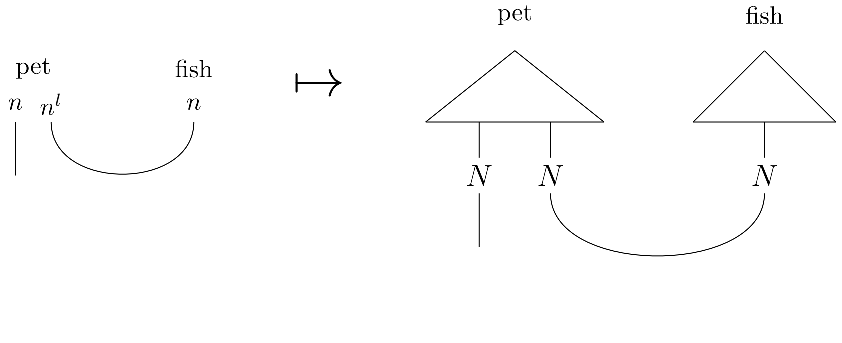
Whenever we have $f_{C_1 \wedge C_2}(x) > min\{f_{C_1}(x), f_{C_2}(x)\}$, we call this *overextension*. This is exactly what we have in the Pet Fish phenomenon, with x = goldfish, C_1 = *pet* and C_2 = *fish*.

Briefly, the problem with applying fuzzy set theory to concepts like this is that is does not allow us to consider the interaction that occurs between the two objects and how they modify each other to form the combined whole. It does not allow us to 'see' what features they have in common. A goldfish is a more typical pet fish than it is a pet or a fish simply because it shares more of the common features to the two concepts than it does with those of the individual words.

4.1.3 Concept combination in a vector-based DisCoCat

The DisCoCat framework allows for interactions and modifications of this type to occur very naturally. In fact, as observed by Lewis and Coecke in [25], in the phrase *pet fish* the word *pet* clearly plays the role of an adjective and should not be treated as a noun.

The compositional model of meaning allows us to take into consideration the grammatical role of the words in the phrase and assign type nn^l to the adjective *pet* and n to the noun *fish*. Then the meaning of the combined concept is given via:



$$(1_N \otimes \varepsilon_N)\left(\overrightarrow{pet\text{-}adj} \otimes \overrightarrow{fish}\right)$$

If we take the adjective to be given by $\overrightarrow{pet\text{-}adj} = \sum_{ij} \alpha_{ij} \overrightarrow{n_i} \otimes \overrightarrow{n_j}$, with respect to the same basis vectors $\{\overrightarrow{n_i}\}_i$ as for the noun $\overrightarrow{fish} = \sum_k \beta_k \overrightarrow{n_k}$, we obtain the meaning of the phrase to be:

$$\begin{aligned}(1_N \otimes \varepsilon_N)\left(\overrightarrow{pet\text{-}adj} \otimes \overrightarrow{fish}\right) &= (1_N \otimes \varepsilon_N)\left(\sum_{ij} \alpha_{ij}\overrightarrow{n_i} \otimes \overrightarrow{n_j} \;\otimes\; \sum_k \beta_k \overrightarrow{n_k}\right)\\ &= \sum_{ijk} \alpha_{ij}\beta_k\, \overrightarrow{n_i}\langle \overrightarrow{n_j} \mid \overrightarrow{n_k}\rangle\\ &= \sum_{ij} \alpha_{ij}\overrightarrow{n_i}\langle \overrightarrow{n_j} \mid \overrightarrow{fish}\rangle.\end{aligned}$$

In fact, the approach taken in [25] is to simplify the model by taking the adjective to be a sum of its attributes, in accordance with [19] and forcing it back into its assigned dimension by using the Frobenius copy Δ operator. In other words, set $\overrightarrow{pet\text{-}adj} = \sum_i \overrightarrow{e_i} = \sum_i \alpha_i^{pet}\overrightarrow{n_i}$ where $\overrightarrow{e_i}$ are the words that co-occur with the adjective, i.e. in this case mostly nouns that are modified by it. Then to return the adjective back to the appropriate dimension, we take:

$$\overrightarrow{pet\text{-}adj}_{copy} = \Delta(\overrightarrow{pet\text{-}adj}) = \sum_i \alpha_i^{pet}\overrightarrow{n_i} \otimes \overrightarrow{n_i}.$$

Then the meaning of *pet fish* becomes:

$$(1_N \otimes \varepsilon_N)\left(\overrightarrow{pet\text{-}adj}_{copy} \otimes \overrightarrow{fish}\right) = \sum_i \alpha_i^{pet}\overrightarrow{n_i}\,\langle \overrightarrow{n_i} \mid \overrightarrow{fish}\rangle = \overrightarrow{pet\text{-}adj} \odot \overrightarrow{fish},$$

where $\odot$ is the pointwise product. So we get that *pet fish* is simply a fish whose each feature is modified by the corresponding feature of the adjective *pet*, where the adjective itself takes into account the arguments that go with it.

The idea then is the following. After computing the vector for the concept *pet fish*, the vector for *goldfish* is compared to it, and also to the vectors for *pet* and *fish* individually. This is done via the symmetric cosine similarity measure.

Definition 23. *Let $\overrightarrow{x} = (x_1, \dots x_n)^T$ and $\overrightarrow{y} = (y_1, \dots, y_n)^T$ be two vectors of the same dimension. Then the cosine between them is given by:*

$$cos\,(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{||\overrightarrow{x}|| \cdot ||\overrightarrow{y}||} = \frac{\sum_i x_i \times y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}.$$

Below is a summary of the toy example performed in [25], exactly as it appears in the paper.

Suppose that the nouns *pet*, *fish*, *goldfish*, *cat*, *dog*, *shark* are modeled with respect to the hand-chosen set of salient features:

$$\{\overrightarrow{cared\text{-}for}, \overrightarrow{vicious}, \overrightarrow{fluffy}, \overrightarrow{scaly}, \overrightarrow{lives\text{-}in\text{-}sea}, \overrightarrow{lives\text{-}in\text{-}house}\},$$

as shown in the table below:

	pet	**fish**	**goldfish**	**cat**	**dog**	**shark**
cared-for	1	0.2	0.7	0.9	0.9	0
vicious	0.2	0.8	0	0.2	0.4	1
fluffy	0.7	0	0	0.9	0.7	0
scaly	0. 2	1	1	0	0	1
lives in sea	0	0.8	0	0	0	1
lives in house	0.9	0.3	0.9	0.9	0.9	0

Take $\overrightarrow{pet\text{-}adj} = \overrightarrow{dog} + \overrightarrow{cat} + \overrightarrow{goldfish} = (0.5, 0.6, 1, 1, 2.7)^T$. Then the cosine similarities between the words are as follows:

	goldfish	**cat**	**dog**	**shark**
pet(noun)	0.7309	0.9816	0.9809	0.1497
fish	0.5989	0.2500	0.3292	0.9670
pet(adj) **fish**	0.9379	0.5550	0.6225	0.5846

Thus, as we can see from the table, the similarity between *goldfish* and *pet fish* is indeed higher than that between *goldfish* and *pet* or *fish* individually.

4.1.4 Transition to a density-matrix based environment in CPM(FHilb)

Following the same train of thought, we introduce the density matrix for the adjective *pet* to be the sum of the density matrices corresponding to the pet nouns that the adjective modifies, i.e. we think of the adjective *pet* as being a mixture of its arguments. The main motivation behind making the switch to a density-matrix based representation is that it allows us to think of the adjective *pet* as a mixture of its constituents, thus making the interactions between their salient features more prominent. It also allows for asymmetric measures of similarity to be applied to the matrices.

Suppose that the density matrices which correspond to pure states for the nouns *cat*, *dog*, *fish* and *goldfish* are given by:

$$\ulcorner dog \urcorner = |dog\rangle\langle dog| = \left(\sum_i \alpha_i \, |n_i\rangle\right)\left(\sum_j \alpha_j \langle n_j|\right) = \sum_{ij} \alpha_i\alpha_j |n_i\rangle\langle n_j|$$

$$\ulcorner cat \urcorner = |cat\rangle\langle cat| = \left(\sum_k \beta_k \, |n_k\rangle\right)\left(\sum_l \beta_l \langle n_l|\right) = \sum_{kl} \beta_k\beta_l \, |n_k\rangle\langle n_l|$$

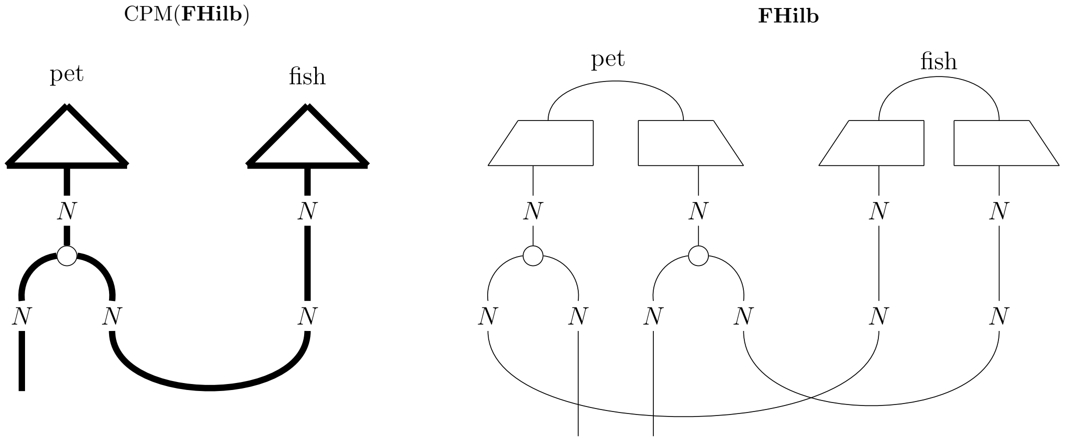
$$\ulcorner goldfish \urcorner = |goldfish\rangle\langle goldfish| = \left(\sum_p \gamma_p \, |n_p\rangle\right)\left(\sum_q \gamma_q \langle n_q|\right) = \sum_{pq} \gamma_p\gamma_q \, |n_p\rangle\langle n_q|$$

$$\ulcorner fish \urcorner = |fish\rangle\langle fish| = \left(\sum_t \delta_t \, |n_t\rangle\right)\left(\sum_u \delta_u \, |n_u\rangle\right) = \sum_{tu} \delta_t\delta_u \, |n_t\rangle\langle n_u|$$

And the mixed state for the adjective *pet* is given by the density matrix:

$$\begin{aligned}\ulcorner \textit{pet-adj} \urcorner &= \ulcorner dog \urcorner + \ulcorner cat \urcorner + \ulcorner goldfish \urcorner \\ &= \sum_{ij} \alpha_i\alpha_j \, |n_i\rangle\langle n_j| + \sum_{kl} \beta_k\beta_l \, |n_k\rangle\langle n_l| + \sum_{pq} \gamma_p\gamma_q \, |n_p\rangle\langle n_q| \\ &= \sum_{rs} C^{pet}_{rs} \, |n_r\rangle\langle n_s|.\end{aligned}$$

As before, in order to compute the meaning of *pet fish*, we will first copy the adjective and then combine it with the noun with the ε-map. Diagrammatically, we get:



The meaing in CPM(**FHilb**) is thus given by:

$$\begin{aligned}
&(1_N \otimes \varepsilon_N) \circ (\Delta_N \otimes 1_N)\, (\ulcorner \textit{pet-adj} \urcorner \otimes \ulcorner \textit{fish} \urcorner) \\
&= (1_N \otimes \varepsilon_N) \circ (\Delta_N \otimes 1_N) \left(\sum_{rs} C^{pet}_{rs} \, |n_r\rangle\langle n_s| \;\otimes\; \sum_{tu} \delta_t \delta_u \, |n_t\rangle\langle n_u| \right) \\
&= (1_N \otimes \varepsilon_N) \left(\sum_{rs} C^{pet}_{rs} \, |n_r\rangle\langle n_s| \otimes |n_r\rangle\langle n_s| \;\otimes\; \sum_{tu} \delta_t \delta_u \, |n_t\rangle\langle n_u| \right) \\
&= \sum_{rstu} C^{pet}_{rs} \, \delta_t \, \delta_u \, |n_r\rangle\langle n_s| \, \langle n_r \,|\, n_t\rangle\langle n_s \,|\, n_u\rangle \\
&= \sum_{rs} C^{pet}_{rs} \, \delta_r \, \delta_s \, |n_r\rangle\langle n_s| \\
&= \ulcorner \textit{pet-adj} \urcorner \odot \ulcorner \textit{fish} \urcorner
\end{aligned}$$

The measure of similarity of density matrices that corresponds closely to the cosine measure for vectors and that will be used here in a similar fashion as above is that of fidelity. The advantage of using fidelity over other symmetric distance measures on density matrices, such as trace distance and trace inner product, lies in the fact that for words represented by pure states in CPM(**FHilb**) the fidelity is equal to the cosine between their corresponding one-dimensional projections, i.e. their **FHilb** counterparts.

Definition 24. *If ρ and σ are two density matrices then the fidelity between them is given by:*

$$F(\rho, \sigma) = Tr\left[\sqrt{\sqrt{\rho}\,\sigma\,\sqrt{\rho}} \right].$$

Fidelity is a symmetric measure of similarity, i.e. we have that $F(\rho, \sigma) = F(\sigma, \rho)$.

If ρ and σ are both pure states, then the fidelity between them is simply $F(\rho, \sigma) = |\langle \varphi \,|\, \psi \rangle|$, where $\rho = |\varphi\rangle\langle\varphi|$ and $\sigma = |\psi\rangle\langle\psi|$. Thus, when comparing the nouns that are represented by pure states via the fidelity measure, we simply recover the results that we had before. For example,

$$F(\ulcorner \textit{fish} \urcorner, \ulcorner \textit{goldfish} \urcorner) = cos(\overrightarrow{fish}, \overrightarrow{goldfish}) = 0.5989$$

The difference is in the fidelity between *pet fish*, represented by $\ulcorner \textit{pet-fish} \urcorner$ and the rest of the nouns. In particular, we get

$$F(\ulcorner \textit{pet-fish} \urcorner, \ulcorner \textit{goldfish} \urcorner) = 0.7934$$

Comparing *pet fish* with the rest of the nouns via the fidelity gives us the following results:

	goldfish	**cat**	**dog**	**shark**
pet(adj) **fish**	0.7934	0.3906	0.4409	0.5136

We observe that all the results obtained are in fact lower than those achieved via computing the cosine between their vectors. In particular, the decrease in similarity between *pet fish* and *goldfish* compared to before seems to match our intuition more closely - a goldfish is indeed a fairly typical pet fish, but a degree of similarity of 0.94 indicated a close relationship between the two bordering on complete overlap, which should not be the case. The density matrix model allows for the shared salient features between the concepts to become more obvious, and at the same time also results in their differences becoming more prominent, and hence the result obtained from the comparison becoming lower.

4.2 Asymmetry of similarity judgments

4.2.1 Similarity is not symmetric

Concept similarity is another fundamental problem in psychology that also manifests itself in cognitive linguistics. The problem with many of the mathematical treatments of similarity is that they are inherently symmetric, especially when based on geometric reasoning. As exhibited in [40] experimentally, concept similarity is intrinsically asymmetric. In the same paper, Tversky argues that similarity judgments should be treated as statements of the form *'a is like b'* (which is usually different from *'b is like a'*), rather than *'a and b are similar'*. He further claims that "the direction of the asymmetry is determined by the relative salience of the stimuli; the variant is more similar to the prototype than vice versa".

In one of the experiments conducted in support of this claim, a group of test subjects is asked to decide which of a given pair of countries is more similar to the other one. The subjects had to chose between *'Country A is similar to country B'* and *'Country B is similar to country A'*. In this experiment almost all of the subjects picked the phrase *'North Korea is similar to China'* and not *'China is similar to North Korea'*. The general observation was that given two countries which share some common features, the more prominent country ('the prototype') determined the direction of the asymmetry.

4.2.2 DisCoCat model to capture asymmetry

The distributional model of meaning framework provides us with a very natural environment for modeling this kind of asymmetry. We propose one way of doing this in **FHilb**. We choose the sentence space S to be one-dimensional truth theoretic and constant, i.e., $\overrightarrow{1}$ everywhere, and model the verb: *is similar to* with respect to the same basis vectors $\{\overrightarrow{n_i}\}_i$ as those used as context words for the noun space N. Thus, the weights C_{ij} for the verb

$$\overrightarrow{is\text{-}similar\text{-}to} = \sum_{ij} C_{ij}\, \overrightarrow{n_i} \otimes \overrightarrow{1} \otimes \overrightarrow{n_j} \;\cong\; \sum_{ij} C_{ij}\, \overrightarrow{n_i} \otimes \overrightarrow{n_j}$$

correspond to the number of times that a noun with salient feature n_i occurs as the subject of the verb, while at the same time a noun with feature n_j appears as its object. In reality, there is no reason why in general we should have $C_{ij} = C_{ji}$, i.e. there is no reason why the verb *is similar to* should be symmetric. This asymmetry of the verb propagates through the sentence via the meaning map and results in an output that differentiates between the two sentences *'China is similar to North Korea'* and *'North Korea is similar to China'*. Note that by essentially eliminating the sentence space, we force the output of the meaning map applied to these sentences to be a real number. Then the idea is that the greater number of the two corresponds to the more 'likely' sentence.

To make this idea concrete, we will consider a toy example. Note that all the weights in this example were created by hand and by intuition and are vaguely based on widely available facts but not extracted from a corpus or supported by any kind of experimental evidence. Empirical evidence would be required to verify the results.

The assumption here is that the countries China and North Korea are judged against a set of context words which correspond to salient features that could in practice be extracted from a corpus or elicited from human trials. We will take the context words to be the following set:

$$\{\overrightarrow{big}, \overrightarrow{populous}, \overrightarrow{prominent}, \overrightarrow{affluent}, \overrightarrow{East\ Asian}, \overrightarrow{communist}, \overrightarrow{militarised}\}.$$

The vectors for $\overrightarrow{China} = \sum_i \alpha_i^{Ch}\, \overrightarrow{n_i}$ and $\overrightarrow{NorthKorea} = \sum_l \alpha_l^{NK}\, \overrightarrow{n_l}$ are summarised in the following table:

	China	**North Korea**
big	0.9	0.3
populous	1	0.4
prominent	0.8	0.4
affluent	0.5	0.2
East Asian	1	1
communist	0.6	0.8
militarised	0.6	0.9

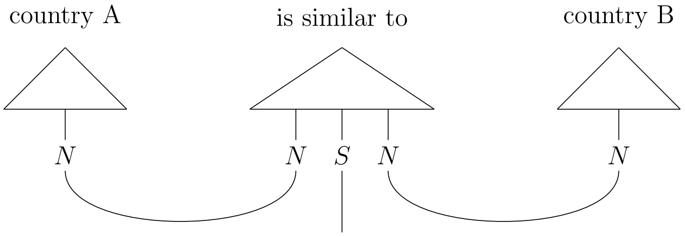
and the weights C_{ij} for the verb $\overrightarrow{is\text{-}similar\text{-}to}$ are given by:

	big	**populous**	**prominent**	**affluent**	**East Asian**	**communist**	**militarised**
big	0.7	0.5	0.6	0.5	0.1	0.2	0.3
populous	0.2	0.7	0.8	0.5	0.6	0.1	0.2
prominent	0.4	0.5	0.7	0.8	0.4	0.1	0.3
affluent	0.6	0.6	0.8	0.6	0.3	0.2	0.2
East Asian	0.3	0.7	0.6	0.3	0.7	0.1	0.4
communist	0.1	0.1	0.2	0.1	0.3	0.8	0.5
militarised	0.2	0.1	0.4	0.2	0.2	0.1	0.5

Next we compute the meanings of the two sentences:

$$\Phi = \text{China is similar to North Korea.}$$
$$\Psi = \text{North Korea is similar to China.}$$

in **FHilb** according to the diagram:



with meaning map $\varepsilon_N \otimes 1_S \otimes \varepsilon_N$, which we can simply treat as $f = \varepsilon_N \otimes \varepsilon_N$ and forget about the sentence space. The meanings are then:

$$\begin{aligned}
f(\Phi) &= (\varepsilon_N \otimes \varepsilon_N)\left(\sum_i \alpha_i^{Ch}\,\overrightarrow{n_i} \;\otimes\; \sum_{jk} C_{jk}\,\overrightarrow{n_j} \otimes \overrightarrow{n_k} \;\otimes\; \sum_l \alpha_l^{NK}\,\overrightarrow{n_l}\right)\\
&= \sum_{ik} \alpha_i^{Ch}\, C_{ik}\, \alpha_k^{NK} \approx 8.1\\
f(\Psi) &= (\varepsilon_N \otimes \varepsilon_N)\left(\sum_l \alpha_l^{Nk}\,\overrightarrow{n_l} \;\otimes\; \sum_{jk} C_{jk}\,\overrightarrow{n_j} \otimes \overrightarrow{n_k} \;\otimes\; \sum_i \alpha_i^{Ch}\,\overrightarrow{n_i}\right)\\
&= \sum_{ij} \alpha_j^{NK}\, C_{ji}\, \alpha_i^{Ch} \approx 9.1
\end{aligned}$$

Thus, as expected, $f(\Psi) > f(\Phi)$.

Discussion

As observed in the above example, this very simple model seems to be applicable to the task of modeling asymmetry and there is no reason not to be believe that it can be extended to all sorts of examples and possible scenarios. What enables us to do this is the intrinsic asymmetry of verbs, even the auxiliary verb *to be*, and the fact that the compositionality of the systems allows for this asymmetry to propagate through the sentence and result in a different meaning outputs to *'Noun1 verb Noun2'* and *'Noun2 verb Noun1'*.

The same idea can easily be extended to the category CPM(**FHilb**) by making the transition from noun vectors to matrices. However, the results obtained would provide only a marginal improvement at the expense of increased complexity.

An alternative solution that eliminates the need to use the verb *is similar to* or any of its synonyms would be to model the two concepts (countries) as mixed states in CPM(**FHilb**) and compare them via an asymmetric measure of similarity on density matrices that produces quantitative results. We will introduce one such measure in the next chapter and discuss towards the end how it could potentially be implemented on the task in future work.

Chapter 5

P-Hyponymy

5.1 Introduction

As we already mentioned, one of the greatest advantages of transitioning to the CPM(**FHilb**) category and using density matrices instead of vectors lies in the possibilities for applying asymmetric measures and orders on the matrices. One such measure was developed in [2] and was utilised for the task of determining word hyponymy. Here we will introduce a new and simpler measure, which not only allows for exhibiting hyponymy relations between concepts, but also enables us to quantify them. We will see how this works on the level of sentences and how it can be generalised to apply to all familiar grammatical structures within the DisCoCat framework. Finally, we will briefly consider what happens when we try to implement variations of this measure to model other phenomena, such as those mentioned in Chapter 4.

5.1.1 What is hyponymy?

In simplest terms possible, hyponymy is a '*is-a-type-of*' relation between two concepts X (the hyponym) and Y (the hypernym), i.e. X and Y are in a hyponym-hypernym relation if *X is a type of Y*. For example, a *Siamese cat* is a type of *cat*. However, in reality, hyponymy is an incredibly complex linguistic phenomenon and has no universally agreed upon rigorous mathematical definition.

Examining the various possible characterisations of hyponymy even on a superficial level is well beyond the scope and intention of this work. The interested reader is referred to [10]. Broadly speaking, there are several different ways of defining this linguistic concept, including various extensional and intentional logic, collocational, componential, and prototype-theoretic approaches. Each of these has certain advantages and disadvantages over the others, and there are inevitably always cases of real-life uses of hyponymy that each of them fails to capture properly, or at all.

The definition that will be assumed here will be more or less a simplified version of the prototype-theoretic one. As we said at the beginning, the hyponym-hypernym relation X-Y is one of the form *X is a type/kind/sort of Y*. To make this a bit more precise, we first need to specify that X and Y can be any pair of concepts expressible via words or phrases of the same grammatical type, and in terms of a pre-defined set of salient features. Then we will say that **X is a hyponym of Y** and that **Y is a hypernym of X** if the features of X are contained in those of Y.

It will be assumed below that whenever we write 'X-Y pair', we mean a pair of concepts where X is a hyponym of Y. Similarly, when we say that 'we have hyponymy' between X and Y, this will be understood to mean that X is a hyponym of Y.

5.1.2 Graded hyponymy

Note that, clearly, some X-Y pairs better exemplify hyponymy than others. For example, the pair *apple - fruit* as opposed to *tomato - fruit*. This leads to another concept, that of **typicality**, or **prototypicality**. Loosely speaking, if the hyponymy bond between X and Y is sufficiently strong, then we can think of them as being in a '*is-a-typical*' kind of relationship, rather than a '*is-a-type-of*' one. On the other hand, if we have very 'weak' hyponymy, then we can think of X as being just a member of the class Y. Thus, while this is not at all rigorous from a linguistic point of view, it will be useful for us to consider membership, hyponymy and typicality as increasingly stronger versions of the same concept, i.e. we will assume that *membership* $\leq$ *hyponymy* $\leq$ *typicality*. In fact, we will unite these into the single new concept that we will call **p-hyponymy**, which will be a graded version of hyponymy that does not necessarily coincide entirely with the linguistic definition of the concept.

We will not only be interested in whether or not X-Y is a hyponym-hypernym pair, but also in how strong the hyponymy between these concepts is. The idea is that the stronger this hyponymy is, the closer we get to having prototypicality. For example, a *shinai* is a hyponym to *kendo weapon*; in fact, a shinai is the prototypical kendo weapon. Hence, we expect to have very high hyponymy between the two, as captured by the value of p. What we mean by that will be made precise below.

5.1.3 Using measures on density matrices for hyponymy

Modeling hyponymy in the DisCoCat framework was first considered by Balkir in [2] where she introduced an asymmetric similarity measure on density matrices based on quantum relative entropy, which can be used to translate hyponym-hypernym relations to the level of positive transitive sentences. This measure relies on a version of the Distributional Inclusion Hypothesis and, while it is possible to make it quantitative, it is only considered in its qualitative version which induces a partial order on density matrices. Our aim here will be to provide an alternative and simpler measure, relying only on the properties of density matrices and the fact that they are the states in CPM(**FHilb**). This will enable us to order the words captured via the density matrices based on the *strength* of their relative hyponymy, i.e. in a quantitative fashion, as described by the idea of *p-hyponymy*. We will show how the order induced by the p-hyponymy can be lifted to the sentence level, not only for positive transitive sentences, but for a much wider range of structures. In a sense, this measure will also be more general than that in [2] as our definition of hyponymy is much more general.

5.1.4 Properties of hyponymy

Before proceeding with defining the concept of *p-hyponymy*, we will list a couple of properties of hyponymy. We will show later that these can be captured by our new measure.

- **Asymmetry.** If X is a hyponym of Y, then this does not imply that Y is a hyponym of X. In fact, we may even assume that only one of these relationships is possible, and that they are mutually exclusive. For example, *tchoukball* is a type of *sport* and hence tchoukball-sport is a hyponym-hypernym pair. However, *sport* is definitely not a type of *tchoukball.*
- **Pseudo-transitivity.** If X is a hyponym of Y and Y is a hyponym of Z, then X is a hyponym of Z. For example, a *Volkswagen* is a type of *car*, and a *car* is a type of *vehicle*, so we have the hyponym-hypernym pairs Volkswagen-car and car-vehicle. Then it is certainly true that a *Volkswagen* is a type of *vehicle.* However, the relationship between hyponyms becomes weaker with the distance between them. A Volkswagen might sill be a vehicle, but is it definitely more of a car, and similarly, the more general concept of a car is closer in meaning to the concept of a vehicle than a Volkswagen is. This is why we call this pseudo transitivity. In a sense, this implies a hierarchical structure of hyponyms, where the further away we go from the lowest hyponym, the weaker the hyponymy. This is a natural consequence of the fact that in general a hypernym defines a broader category than its hyponyms.

5.2 Background definitions and results

Here we present some definitions and results that will be used later on.

Definition 25 (Positive semi-definite matrix)**.** *Let $M \in M_n(\mathbb{R})$ be a real symmetric $n \times n$ matrix. We say that M is positive semi-definite, and write $M \succeq 0$, if for any column vector $\overrightarrow{x} \in \mathbb{R}^n$ it holds* *that:*

$$\overrightarrow{x}^T M \overrightarrow{x} \geq 0.$$

Note that $\overrightarrow{x}^T M \overrightarrow{x} = \langle \overrightarrow{x}, M\overrightarrow{x} \rangle$, where $\langle \bullet, \bullet \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is the usual vector inner product on $\mathbb{R}^n$.

- An alternative characterisation of positive semi-definiteness in terms of eigenvalues tells us that a square Hermitian matrix M is positive semi-definite iff all of its eigenvalues are non-negative.
- Note that density matrices can be characterised in terms of positive semi-definite matrices as density matrices are self-adjoint, positive semi-definite matrices with trace 1.

Proposition 1. *The sum of an arbitrary number of positive scalar multiples of positive semi-definite matrices is positive semi-definite.*

Proof. Fix $n \in \mathbb{N}$ and let $\mathbf{X_1}, \ldots, \mathbf{X_n}$ be positive semi-definite matrices and $\alpha_1, \ldots \alpha_n$ be non-negative real numbers. Let $\overrightarrow{x} \in \mathbb{R}^n$ be an arbitrary non-zero vector and consider the inner product $\langle \overrightarrow{x}\ , \sum_i \alpha_i \mathbf{X_i} \cdot \overrightarrow{x} \rangle$. By the linearity of the inner product, we have that:

$$\left\langle \overrightarrow{x}\ , \sum_i \alpha_i \mathbf{X_i} \cdot \overrightarrow{x} \right\rangle = \sum_i \alpha_i \left\langle \overrightarrow{x}, \mathbf{X_i} \cdot \overrightarrow{x} \right\rangle .$$

Since for each $j \in [1, n]$, $\mathbf{X_j}$ is a positive semi-definite matrix, we have that each $\langle \overrightarrow{x}\ , \mathbf{X_j} \cdot \overrightarrow{x} \rangle \geq 0$ and hence $\sum_i \alpha_i \langle \overrightarrow{x}\ , \mathbf{X_i} \cdot \overrightarrow{x} \rangle \geq 0$. □

Proposition 2. *The following is a necessary condition for a Hermitian matrix $A = (a_{ij})$ to be positive semi-definite: $a_{ii} \geq 0$, $\forall i$, i.e. all the diagonal entries are non-negative.*

5.3 P-Hyponyms

5.3.1 A new measure on density matrices

The measure of hyponymy that we described above and named *p-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. This is based on the partial order on the set of all square matrices given by $A \leq B$ iff $B - A \succeq 0$.

Definition 26 (P-hyponym)**.** *Let $\ulcorner A \urcorner$ and $\ulcorner B \urcorner$ be density matrix representations of the concepts A and B respectively. We say that A is a* p-hyponym *of B for a given value of* p *in the range* $(0, 1]$ *and write $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner$ if*

$$\ulcorner B \urcorner - p \ulcorner A \urcorner \succeq 0.$$

Remark. Note that such a p need not be unique or even exist at all. We will consider the interpretation and implications of this later on. Moreover, whenever we do have p-hyponymy between A and B, there is necessarily a largest such p.

Definition 27 (P-max hyponym)**.** *If A is a p-hyponym of B for any $p \in (0, 1]$, then there is necessarily a maximal possible such p. We denote it by p_{max} and define it to be the max value of p in the range* $(0, 1]$ *for which we have $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner$, in the sense that there does not exist $q \in (0, 1]$ s.t. $q > p$ and $\ulcorner A \urcorner \preccurlyeq_q \ulcorner B \urcorner$.*

5.3.2 Interpretation of the values of *p*

The values of p in the p-hyponymy measure are meant to denote probabilities. The idea is that the closer p is to 1, the stronger the hyponymy between A and B. That is to say, the closer A is to being in a '*is-a-typical*' relationship with B. Clearly, the value of p is not unique for an A-B pair; however, p-max is always unique. This can be interpreted as giving us an upper bound of how close A can get to B, while at the same time indicating that it might not always be the case that A does get that close to B. This might, for example, depend on the context in which the two concepts are used. A *shinai* may be a very strong p-max hyponym of *kendo weapons* in general, but exhibit a weaker connection to the martial art in some context.

Below we give a brief justification of why the range of possible values of p makes sense.

Proposition 3 (P-hyponymy for $p \leq 0$ is useless)**.** *It always holds that* $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner$ *if* $p \leq 0$. *Therefore, the notion of* p*-hyponymy is useless for non-positive values of* p.

Proof. Let $p \leq 0$. Then

$$\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner \iff \ulcorner B \urcorner - p \ulcorner A \urcorner \succeq 0 \iff \ulcorner B \urcorner + q \ulcorner A \urcorner \succeq 0\,,$$

where $q = -p \geq 0$. Since $\ulcorner A \urcorner$ and $\ulcorner B \urcorner$ are density matrices and hence positive semi-definite, and q is non-negative, we get by **Proposition 1** that $\ulcorner B \urcorner + q \ulcorner A \urcorner \succeq 0$. □

The next proposition justifies why we do not consider values of p exceeding 1. It turns out that we never actually find ourselves in this situation. This makes sense intuitively, given our interpretation of the value of p as being a measure of proximity or a probability. If $p = 1$ is interpreted as being absolute hyponymy, or prototypicality, then we should expect to not be able to exceed this value.

Proposition 4. *The value of* p *cannot exceed 1.*

Note that we assume that the matrices we are comparing are always of the same dimension. Also, since the density matrices correspond to words or phrases in our model, we know that all of their entries are non-negative, given the methods by which word meaning vectors are normally constructed.

Proof. Suppose that p-hyponymy were possible for values of p exceeding 1. Let p be such a value for which we have $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner$, for some density matrices $\ulcorner A \urcorner$ and $\ulcorner B \urcorner$. Then $\ulcorner B \urcorner - p \ulcorner A \urcorner \succeq 0$. Define the matrix $C = \ulcorner B \urcorner - p \ulcorner A \urcorner$. By assumption, this matrix is positive semi-definite and has diagonal entries $c_{ii} = b_{ii} - pa_{ii}$. Since a positive semi-definite matrix has only non-negative entries on its diagonal, by **Proposition 2**, we have that $c_{ii} \geq 0\,, \forall i$. Also, the matrices $\ulcorner A \urcorner$ and $\ulcorner B \urcorner$ have only non-negative entries, and in particular $a_{ii} \geq 0\,, \forall i$ and $b_{ii} \geq 0\,, \forall i$. Moreover, since they are density matrices, we have $\sum_i b_{ii} = \sum_j a_{jj} = 1$. Thus,

$$(c_{ii} \geq 0 \ \forall i) \implies (b_{ii} \geq pa_{ii} \ \forall i) \implies \sum_i b_{ii} \geq p \sum_j a_{jj} > \sum_j a_{jj} \implies 1 > 1\,,$$

which is a contradiction. □

The following observation relates to the property of words to be hyponyms of themselves in a trivial way. In other words, we expect that any word or concept should be an absolute hyponym of itself, in the sense of being a 1-max hyponym of itself.

Proposition 5. *Any word A is a 1-max hyponym of itself.*

Proof. Let A be an arbitrary word with density matrix representation $\ulcorner A \urcorner$. Then as $\ulcorner A \urcorner - \ulcorner A \urcorner = \ulcorner 0 \urcorner$ and $\ulcorner 0 \urcorner$ is by definition a positive semi-definite matrix, we conclude that $\ulcorner A \urcorner \preccurlyeq_1 \ulcorner A \urcorner$, i.e. A is a 1-hyponym of itself, and hence a 1-max hyponym of itself. □

5.3.3 Extracting p-hyponymy values out of hyponym-hypernym pairs

We will make the assumption that hypernyms can be expressed in terms of their hyponyms, which is, again, not completely rigorous from a linguistic point of view, but is a valid assumption for our purposes nonetheless. For example, if all the hyponyms of the hypernym *sport* are *tchoukball*, *volleyball* and *pickleball*, then we can think of the concept *sport* as being *tchoukball*+*volleyball*+*pickleball*.

In practice, one way of making use of this assumption is by representing the density matrix corresponding to *sport* as a mixture of the density matrices for the individual sports, weighted by the number of times that the given sport has co-occurred with the word *sport* in a large body of text or corpus. This idea is somewhat similar to the way in which we sometimes represent collective nouns as a sum of their constituents in the vector-based model. The weights are generally normalised to be between 0 and 1 and are not assumed to sum to 1. However, it will be more convenient for us to assume that they do, i.e. to treat them as probabilities. We can always get rid of this assumption by normalising the resulting matrix for *sport* so that it has trace 1 after we have already included all the sports in it.

More generally, this works as follows. Suppose that $\ulcorner X_i \urcorner$ are the density matrix representations for the hyponyms of the word B. Then the density matrix for B is given by:

$$\ulcorner B \urcorner = \sum_i p_i \ulcorner X_i \urcorner.$$

Here $\sum_i p_i = 1$. To drop this assumption, we can define $\ulcorner B' \urcorner = \sum_i q_i \ulcorner X_i \urcorner$ to be the normalised version of $\ulcorner B \urcorner$ with trace 1. We will always assume any density matrix we have is already normalised.

Proposition 6. *Suppose that* $\ulcorner A \urcorner \in \{\ulcorner X_i \urcorner\}_i$*, i.e that we have:*

$$\ulcorner B \urcorner = p_j \ulcorner A \urcorner + \sum_{i \neq j} p_i \ulcorner X_i \urcorner$$

Then

$$\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner,$$

for any $p \leq p_j$*.*

Proof. Without loss of generality suppose that $\ulcorner B \urcorner = p_1 \ulcorner A \urcorner + \sum_{i \neq 1} p_i \ulcorner X_i \urcorner$ and consider

$$\xi = \langle \overrightarrow{x}, (\ulcorner B \urcorner - p \ulcorner A \urcorner) \overrightarrow{x} \rangle \quad \text{for any} \quad p \leq p_1.$$

We want to show that $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner$, i.e. $\xi \geq 0$, $\forall \overrightarrow{x} \neq 0$. We have:

$$\begin{aligned}
\xi = \langle \overrightarrow{x}, (\ulcorner B \urcorner - p \ulcorner A \urcorner) \overrightarrow{x} \rangle &= \left\langle \overrightarrow{x}, \left((p_1 \ulcorner A \urcorner + \sum_{i \neq 1} p_i \ulcorner X_i \urcorner) - p \ulcorner A \urcorner \right) \overrightarrow{x} \right\rangle \\
&= \left\langle \overrightarrow{x}, \left((p_1 - p) \ulcorner A \urcorner + \sum_{i \neq 1} p_i \ulcorner X_i \urcorner \right) \overrightarrow{x} \right\rangle = \left\langle \overrightarrow{x}, (p_1 - p) \ulcorner A \urcorner \overrightarrow{x} + \sum_{i \neq 1} p_i \ulcorner X_i \urcorner \overrightarrow{x} \right\rangle \\
&= \langle \overrightarrow{x}, (p_1 - p) \ulcorner A \urcorner \overrightarrow{x} \rangle + \left\langle \overrightarrow{x}, \sum_{i \neq 1} p_i \ulcorner X_i \urcorner \overrightarrow{x} \right\rangle \\
&= (p_1 - p) \langle \overrightarrow{x}, \ulcorner A \urcorner \overrightarrow{x} \rangle + \sum_{i \neq 1} p_i \ \langle \overrightarrow{x}, \ulcorner X_i \urcorner \overrightarrow{x} \rangle
\end{aligned}$$

By assumption, all of the $\ulcorner X_i \urcorner$ and $\ulcorner A \urcorner$ are positive semi-definite matrices and all the p_i are non-negative. For any choice of p s.t. $p \leq p_1$ we have $p_1 - p \geq 0$. Thus, for any such p we end up with a non-negative linear combination of non-negative quantities, i.e. a non-negative quantity. In other words, we have p-hyponymy. □

The p-max value: Discussion and assumptions

From the above proof we notice that the value p_1 definitely gives us p_1-hyponymy between A and B, but it is actually possible that there exists a value, say q, such that $q > p_1$ and for which we have q-hyponymy between A and B. Indeed, this happens whenever we have a q for which $(p_1 - q) \langle \overrightarrow{x}, \ulcorner A \urcorner \overrightarrow{x} \rangle \geq -\sum_{i \neq 1} p_i \ \langle \overrightarrow{x}, \ulcorner X_i \urcorner \overrightarrow{x} \rangle$. Thus, p_1 may not be the maximum value for hyponymy between A to B. In practice, such a value can be determined by testing for numbers $q \in (0, 1]$, s.t. $q > p_1$ and for which the eigenvalues of the matrix $\ulcorner B \urcorner - q \ulcorner A \urcorner$ are all non-negative. This follows from the equivalent characterisation of positive semi-definiteness in terms of eigenvalues. If $\ulcorner A \urcorner$ is not in the span of the rest of the $\ulcorner X_i \urcorner$ then we do get that the maximum value of the hyponymy is $p = p_1$. We will work under the assumption that this is always the case, i.e. that the co-hyponyms of a hypernym are independent of each other.

Now suppose that we drop the assumption that we have previous knowledge about all the hyponyms of B that are meant to go into the mixture that we use to define B, i.e. suppose that we do not know that $\ulcorner B \urcorner = \sum_i p_i \ulcorner X_i \urcorner$, where $\ulcorner X_i \urcorner$ are all the relevant hyponyms of B. Suppose that all we know is that A *is* a hyponym of B. Can we still say anything about the actual strength of the hyponymy between A and B?

According to our definition, starting only from the knowledge that A is a hyponym of B, we get that $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner$, for some p, and hence that $\ulcorner B \urcorner - p \ulcorner A \urcorner \succeq 0$. So there exists some PSD matrix, say ρ, such that $\ulcorner B \urcorner = p \ulcorner A \urcorner + \rho$. Then, similar to above, to determine the maximum possible value of p, we will need to test for values in the range $(0, 1]$ for which the eigenvalues of the matrix $\ulcorner B \urcorner - p \ulcorner A \urcorner$ are all non-negative. So, in theory, we could find a maximal p which satisfies the definition of p-hyponymy between A and B. However, at present we have no way of knowing straight away whether this value is at all useful to us. In order to find out, a large-scale practical experiment would need to be carried out to obtain empirical data that can give us an indication of the validity of this model.

Thus, in our applications and examples below, we will always work under the assumption that hypernyms are expressed in terms of a relevant set of independent hyponyms, in which case we have:

$$\ulcorner A \urcorner \preccurlyeq_{p_{max}} \ulcorner B \urcorner \iff \ulcorner B \urcorner = p_{max} \ulcorner A \urcorner + \sum_i p_i \ulcorner X_i \urcorner,$$

as in the previous proposition. Of course, all the results will also work for values of hyponymy below the maximal, but since the p-max hyponymy is what we are primarily interested in, we will assume that we mean p-max hyponymy whenever we say p-hyponymy.

Note that all of the proofs below will also work even if we do not make the above assumption, but rather only work with the original definition that $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner \iff \ulcorner B \urcorner - p \ulcorner A \urcorner \succeq 0 \iff \ulcorner B \urcorner = p \ulcorner A \urcorner + \rho$ for an unknown positive operator ρ. This is because the idea behind all of the proofs is that the morphisms of the category CPM(**FHilb**) are positive operators and hence preserve density matrices.

5.3.4 Properties of P-Hyponymy

The p-hyponymy measure satisfies some of the key properties of hyponymy, as described in the first section.

Property 1 (P-hyponymy is not symmetric)**.** *Given our assumption that hypernyms are expressed in terms of their hyponyms, the asymmetry of the hyponymy relation is satisfied by default.*

Note, however, that we are also quantifying the hyponymy between the words and so it is possible to have both $A \preccurlyeq_p B$ and $B \preccurlyeq_q A$. In this case, if, say, p has a high value then q is bound to be very close to 0, and hence indicative of basically non-existent hyponymy. It is also generally not true that

$$(\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner \text{ and } \ulcorner B \urcorner \preccurlyeq_p \ulcorner A \urcorner) \implies \ulcorner A \urcorner = \ulcorner B \urcorner$$

unless $p = q = 1$, i.e. $A \preccurlyeq_1 B$ and $B \preccurlyeq_1 A$. Then we just have the standard partial order on square matrices induced by the positive semi-definiteness and this implies that in this case we must have $A = B$. In practice, this means that either A and B are the same word or completely overlapping synonyms, assuming that such a thing is even possible.

This is exactly how we expect hyponyms to behave in real life. As an example, we expect *pork* to be a p-hyponym of *meat* for a relatively high value of *p*, and *meat* to be a q-hyponym of *pork* for a very small value of *q*. We certainly do not expect this to imply an equivalence between pork and meat.

A key property of hyponymy in the linguistic sense is that its strength decreases with distance. What we mean by that is that if concept A is a hyponym of concept B and B is a hyponym of C, then A is still a hyponym of C, but weaker than the previous two, like in our vehicle-car-Volkswagen example.

Property 2. *(P-hyponymy decreases with distance) Suppose that A is a p-hyponym of B and B is a p'-hyponym of C, i.e. $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner \preccurlyeq_{p'} \ulcorner C \urcorner$. Then A is a $p'' - hyponym$ of C for $p'' = p \cdot p'$, i.e. $\ulcorner A \urcorner \preccurlyeq_{p''} \ulcorner C \urcorner$.*

Proof. From $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner \preccurlyeq_{p'} \ulcorner C \urcorner$ we get:

$$\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner \iff \langle \overrightarrow{x}, (\ulcorner B \urcorner - p \ulcorner A \urcorner) \overrightarrow{x} \rangle \geq 0 \iff \langle \overrightarrow{x}, \ulcorner B \urcorner \overrightarrow{x} \rangle \geq p \langle \overrightarrow{x}, \ulcorner A \urcorner \overrightarrow{x} \rangle,$$
$$\ulcorner B \urcorner \preccurlyeq_{p'} \ulcorner C \urcorner \iff \langle \overrightarrow{x}, (\ulcorner C \urcorner - p' \ulcorner B \urcorner) \overrightarrow{x} \rangle \geq 0 \iff \langle \overrightarrow{x}, \ulcorner C \urcorner \overrightarrow{x} \rangle \geq p' \langle \overrightarrow{x}, \ulcorner B \urcorner \overrightarrow{x} \rangle.$$

Combining these, we get:

$$\langle \overrightarrow{x}, \ulcorner C \urcorner \overrightarrow{x} \rangle \geq p' \langle \overrightarrow{x}, \ulcorner B \urcorner \overrightarrow{x} \rangle \geq p'p \langle \overrightarrow{x}, \ulcorner A \urcorner \overrightarrow{x} \rangle = p'' \langle \overrightarrow{x}, \ulcorner A \urcorner \overrightarrow{x} \rangle \iff \ulcorner A \urcorner \preccurlyeq_{p''} \ulcorner C \urcorner.$$

□

To illustrate this property, consider the following example.

Example

Let $\{|e_1\rangle, |e_2\rangle, |e_3\rangle, |e_4\rangle\}$ be the standard orthonormal basis for $\mathbb{R}^4$ and consider the words:

$$\overrightarrow{\textit{white chocolate}} = \sum_i \alpha_i \, |n_i\rangle, \quad \overrightarrow{\textit{milk chocolate}} = \sum_j \beta_j \, |n_j\rangle$$
$$\overrightarrow{\textit{dark chocolate}} = \sum_k \gamma_k \, |n_k\rangle, \quad \overrightarrow{\textit{cake}} = \sum_l \delta_l \, |n_l\rangle$$

We take the basis vectors to correspond to the context words *sweet*, *milky*, *spongy* and *high-calorie* and intuitively give the following vector representations to our words, where each word is evaluated against each property on a scale of 0 to 1.

	white chocolate	milk chocolate	dark chocolate	cake
sweet	0.8	0.8	0.5	0.6
milky	0.7	0.5	0.3	0.1
spongy	0	0	0	0.8
high-calorie	0.7	0.6	0.8	0.5

We take the hypernyms *chocolate* and *sweets* to be given by:

$$\ulcorner \textit{chocolate} \urcorner = \frac{1}{4} \ulcorner \textit{white chocolate} \urcorner + \frac{1}{2} \ulcorner \textit{milk chocolate} \urcorner + \frac{1}{4} \ulcorner \textit{dark chocolate} \urcorner$$
$$\ulcorner \textit{sweets} \urcorner = \frac{1}{5} \ulcorner \textit{chocolate} \urcorner + \frac{4}{5} \ulcorner \textit{cake} \urcorner$$

The first hyponymy that we are interested in is that of white chocolate to chocolate and the second one is that of chocolate to sweets. An explicit calculation shows us that the maximum p for which we have:

$$\ulcorner \textit{white chocolate} \urcorner \preccurlyeq_p \ulcorner \textit{chocolate} \urcorner$$
$$\begin{pmatrix} 0.6400 & 0.5600 & 0 & 0.5600 \\ 0.5600 & 0.4900 & 0 & 0.4900 \\ 0 & 0 & 0 & 0 \\ 0.5600 & 0.4900 & 0 & 0.4900 \end{pmatrix} \preccurlyeq_p \begin{pmatrix} 0.5425 & 0.3775 & 0 & 0.4800 \\ 0.3775 & 0.2700 & 0 & 0.3325 \\ 0 & 0 & 0 & 0 \\ 0.4800 & 0.3325 & 0 & 0.4625 \end{pmatrix}$$

is exactly $p = 0.25$.
Similarly, the largest p' for which we get:

$$\ulcorner \textit{chocolate} \urcorner \preccurlyeq_{p'} \ulcorner \textit{sweets} \urcorner$$
$$\begin{pmatrix} 0.5425 & 0.3775 & 0 & 0.4800 \\ 0.3775 & 0.2700 & 0 & 0.3325 \\ 0 & 0 & 0 & 0 \\ 0.4800 & 0.3325 & 0 & 0.4625 \end{pmatrix} \preccurlyeq_{p'} \begin{pmatrix} 0.3965 & 0.1235 & 0.3840 & 0.3360 \\ 0.1235 & 0.0620 & 0.0640 & 0.1065 \\ 0.3840 & 0.0640 & 0.5120 & 0.3200 \\ 0.3360 & 0.1065 & 0.3200 & 0.2925 \end{pmatrix}$$

is $p' = 0.2$. Thus, we expect that the maximum value p'' for which we have:

$$\ulcorner \textit{white chocolate} \urcorner \preccurlyeq_{p''} \ulcorner \textit{sweets} \urcorner$$

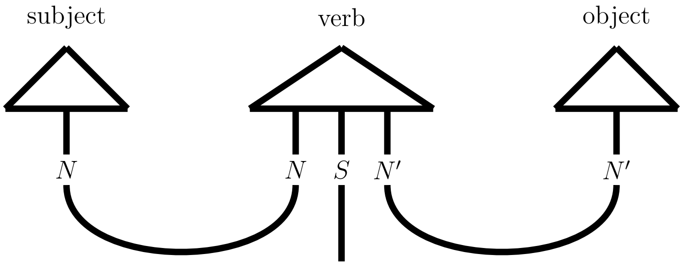
to be $p'' = 0.25 \times 0.2 = 0.05$. Again, a straightforward direct calculation shows that this is indeed the case.

5.4 P-Hyponymy lifted to the level of sentences

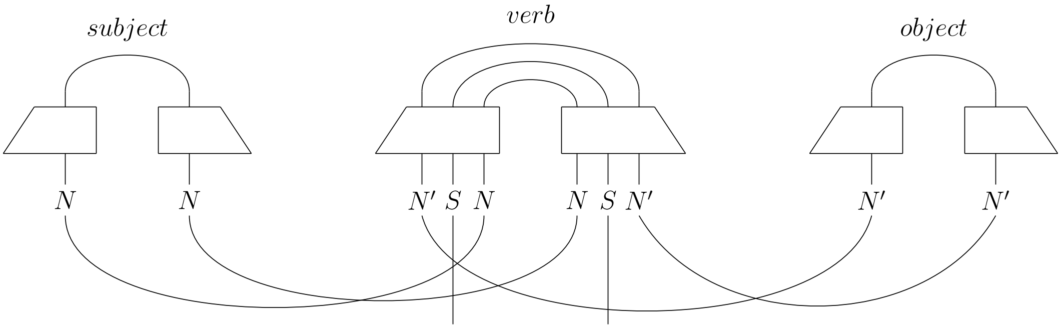
We will now consider what happens when we have two sentences such that one of them contains one or more hyponyms of one or more words from the other. We will show that in this case the hyponymy is 'lifted' to the sentence level, and even the p-values are preserved in a very intuitive fashion. After considering a couple of specific sentence construction, we will generalise this result to account for a broad category of sentence patterns that work in the distributional compositional model.

5.4.1 P-Hyponymy in positive transitive sentences

Recall that a positive transitive sentence has the following diagrammatic representation in CPM(**FHilb**):



Translated to **FHilb**:



So the meaning of the sentence **Subject verb object** is given by:

$$(\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\left(\ulcorner subj \urcorner \otimes \ulcorner verb \urcorner \otimes \ulcorner obj \urcorner\right) \ ,$$

where the epsilon and identity morphisms are those from CPM(**FHilb**). We will represent the subject and object vectors in **FHilb** by:

$$\overrightarrow{subj} = \sum_i \alpha_i^{subj}\, \overrightarrow{n_i} \qquad \text{and} \qquad \overrightarrow{obj} = \sum_j \beta_j^{obj}\, \overrightarrow{n_j}.$$

Their corresponding density matrix representations are given by:

$$\ulcorner subj \urcorner = \sum_{ik} \alpha_i^{subj} \alpha_k^{subj}\, |n_i\rangle\langle n_k| \qquad \text{and} \qquad \ulcorner obj \urcorner = \sum_{jl} \beta_j^{obj} \beta_l^{obj} |n_j\rangle\langle n_l|.$$

Finally, let the verb be given by:

$$\overline{verb} = \sum_{rs} C_{rs}^{verb} |n_r\rangle|s\rangle|n_s\rangle.$$

Its density matrix is:

$$\ulcorner verb \urcorner = \left(\sum_{rs} C_{rs}^{verb}|n_r\rangle|s\rangle|n_s\rangle\right)\left(\sum_{pq} C_{pq}^{verb}\langle n_p|\langle s'|\langle n_q|\right) = \sum_{rspq} C_{rs}^{verb} C_{pq}^{verb}\, |n_r\rangle\langle n_p|\otimes|s\rangle\langle s'|\otimes|n_s\rangle\langle n_q|$$

Relationship between sentences of the type '*A verb C*' and '*B verb D*', where $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner$ and $\ulcorner C \urcorner \preccurlyeq_q \ulcorner D \urcorner$

We will assume that the sentence space, i.e. the vector space corresponding to S, is not truth-theoretic and that hypernyms are always represented in terms of their hyponyms, as before.

Theorem 1. *Let A, B, C and D be nouns with corresponding density matrix representations $\ulcorner A \urcorner$, $\ulcorner B \urcorner$, $\ulcorner C \urcorner$ and $\ulcorner D \urcorner$, such that A is p-hyponym of B and C is a q-hyponym of D, in the sense that:*

$$\ulcorner B \urcorner = p \ulcorner A \urcorner + \sum_i p_i \ulcorner X_i \urcorner \quad \textit{and} \quad \ulcorner D \urcorner = q \ulcorner C \urcorner + \sum_j q_j \ulcorner Y_j \urcorner .$$

Then we have that:

$$\varphi\left(A\, verb\, C\right) \preccurlyeq_{pq} \varphi\left(B\, verb\, D\right),$$

where $\varphi = \varepsilon_N \otimes 1_S \otimes \varepsilon_N$ is the sentence meaning map for positive transitive sentences.

Proof. Let the density matrix corresponding to the verb be given by $\ulcorner Z \urcorner$. Then we can write the meanings of the two sentences as:

$$\varphi\left(A \ verb \ C\right) = \varphi\left(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner\right) = \left(\varepsilon_N \otimes 1_S \otimes \varepsilon_N\right)\left(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner\right)$$

$$\varphi\left(B \ verb \ D\right) = \varphi\left(\ulcorner B \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner D \urcorner\right) = \left(\varepsilon_N \otimes 1_S \otimes \varepsilon_N\right)\left(\ulcorner B \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner D \urcorner\right)$$

Substituting $\ulcorner B \urcorner = p \ulcorner A \urcorner + \sum_i p_i \ulcorner X_i \urcorner$ and $\ulcorner D \urcorner = q \ulcorner C \urcorner + \sum_j q_j \ulcorner Y_j \urcorner$ in the expression for the meaning of *B verb D*, we get:

$$\begin{aligned}
\varphi(\ulcorner B \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner D \urcorner) &= \left(\varepsilon_N \otimes 1_S \otimes \varepsilon_N\right)\left(\ulcorner B \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner D \urcorner\right) \\
&= \varphi\left(\left(p \ulcorner A \urcorner + \sum_i p_i \ulcorner X_i \urcorner\right) \otimes \ulcorner Z \urcorner \otimes \left(q \ulcorner C \urcorner + \sum_j q_j \ulcorner Y_j \urcorner\right)\right) \\
&= \varphi\left(pq\left(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner\right) + p\left(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \sum_j q_j \ulcorner Y_j \urcorner\right)\right. \\
&\qquad \left. + \sum_i p_i \ulcorner X_i \urcorner \otimes \ulcorner Z \urcorner \otimes \left(\sum_j q_j \ulcorner Y_j \urcorner + q \ulcorner C \urcorner\right)\right) \\
&= \varphi\left(p\left(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \sum_j q_j \ulcorner Y_j \urcorner\right) + \sum_i p_i \ulcorner X_i \urcorner \otimes \ulcorner Z \urcorner \otimes \left(\sum_j q_j \ulcorner Y_j \urcorner + q \ulcorner C \urcorner\right)\right) \\
&\qquad + pq\, \varphi\left(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner\right)
\end{aligned} \tag{5.1}$$

Consider $\varphi(\ulcorner B \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner D \urcorner) - pq\, \varphi(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner)$. We get:

$$\begin{aligned}
&\varphi\left(p\left(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \sum_j q_j \ulcorner Y_j \urcorner\right) + \sum_i p_i \ulcorner X_i \urcorner \otimes \ulcorner Z \urcorner \otimes \left(\sum_j q_j \ulcorner Y_j \urcorner + q \ulcorner C \urcorner\right)\right) \\
&= p \sum_j q_j\, \varphi(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner Y_j \urcorner) + \sum_i \sum_j p_i q_j \varphi(\ulcorner X_i \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner Y_j \urcorner) + q \sum_i p_i\, \varphi(\ulcorner X_i \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner)
\end{aligned} \tag{5.2}$$

Since $\ulcorner X_i \urcorner$, $\ulcorner Y_j \urcorner$, $\ulcorner Z \urcorner$, $\ulcorner A \urcorner$ and $\ulcorner C \urcorner$ are all density matrices, all of the following are also density matrices: $\varphi(\ulcorner X_i \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner Y_j \urcorner)$, $\varphi(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner Y_j \urcorner)$, $\varphi(\ulcorner X_i \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner)$. This is because we are working in the CPM(**FHilb**) category and the meaning map φ is a completely positive map, which means that it sends density matrices to density matrices. Moreover, all of the scalars p_i, q_j, p, q are non-negative. Thus, (5.2) is a sum of non-negative scalar multiples of positive semi-definite matrices, and as such is positive semi-definite itself, by **Proposition 1**. We conclude that:

$$\varphi(\ulcorner B \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner D \urcorner) - pq\, \varphi(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner) \succeq 0,$$

and so

$$\varphi(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner) \preccurlyeq_{pq} \varphi(\ulcorner B \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner D \urcorner).$$

□

Now suppose that all we know is that $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner$ and $\ulcorner C \urcorner \preccurlyeq_q \ulcorner D \urcorner$. Then we can write

$$\ulcorner B \urcorner = p \ulcorner A \urcorner + \lambda \text{ and } \ulcorner D \urcorner = q \ulcorner C \urcorner + \mu,$$

for positive operators λ and μ. Then the exact same proof with λ in place of $\sum_i p_i \ulcorner X_i \urcorner$ and μ in place of $\sum_j q_j \ulcorner Y_j \urcorner$ tells us that $\varphi\left(A\, verb\, C\right) \preccurlyeq_{pq} \varphi\left(B\, verb\, D\right)$. In particular, this applies to the case where $p = p_{max}$ and $q = q_{max}$ are the maximum hyponymy values for the two hyponym-hypernym pairs.

Two special cases of the above result occur when we take either the subjects or the objects of the two sentences to be the same, i.e $\ulcorner A \urcorner = \ulcorner B \urcorner$ or $\ulcorner C \urcorner = \ulcorner D \urcorner$.

Corollary 1. *Let A, B, C be nouns with corresponding density matrix representations $\ulcorner A \urcorner$, $\ulcorner B \urcorner$, $\ulcorner C \urcorner$ and such that $\ulcorner B \urcorner = p \ulcorner A \urcorner + \sum_i p_i \ulcorner X_i \urcorner$. Then we have that:*

$$\varphi\left(A\, verb\, C\right) \preccurlyeq_p \varphi\left(B\, verb\, C\right) \;,$$

where $\varphi = \varepsilon_N \otimes 1_S \otimes \varepsilon_N$ is the sentence meaning map.

Proof. This is just a special case of our theorem with $\ulcorner C \urcorner = \ulcorner D \urcorner$ and $q = 1$. Then (5.2) above becomes simply:

$$\varphi(\ulcorner B \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner) - p\, \varphi(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner) = \sum_i p_i\, \varphi(\ulcorner X_i \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner) \;,$$

which is a positive semi-definite matrix, and thus $\varphi(\ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner) \preccurlyeq_p \varphi(\ulcorner B \urcorner \otimes \ulcorner Z \urcorner \otimes \ulcorner C \urcorner)$. □

Corollary 2. *Let A, C, D be nouns with corresponding density matrix representations $\ulcorner A \urcorner$, $\ulcorner C \urcorner$, $\ulcorner D \urcorner$ and such that $\ulcorner D \urcorner = p \ulcorner C \urcorner + \sum_j q_j \ulcorner Y_j \urcorner$. Then we have that:*

$$\varphi\left(A\, verb\, C\right) \preccurlyeq_p \varphi\left(A\, verb\, D\right) \;,$$

where $\varphi = \varepsilon_N \otimes 1_S \otimes \varepsilon_N$ is the sentence meaning map.

Proof. Similar to above. □

Again, these results also work if we just have that $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner$ or $\ulcorner C \urcorner \preccurlyeq_p \ulcorner D \urcorner$ without any further assumption of knowledge about the representation of the hypernyms.

Examples of p-hyponymy in positive transitive sentences

Example 1

We assumed that the sentence space S is not truth-theoretic. The following example illustrates what happens to positive transitive sentence hyponymy if we take a truth-theoretic approach to sentence meaning, i.e. if we take the sentence space to be one- or two-dimensional truth-theoretic.

Suppose that our sentence space S is 1-dimensional, with its single non-trivial vector being $\overrightarrow{1}$. We will take $\overrightarrow{1}$ to stand for *True* and $\overrightarrow{0}$ for *False*. The sentences we will consider are:

A := Annie likes holidays.
B := Students like holidays.

Let the vector space for the subjects of the sentences be $\mathbb{R}^3 = Span_{\mathbb{R}}\{\left(\begin{smallmatrix}1\\0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}0\\1\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right)\}$, where:

$$\overrightarrow{Annie} = |e_1\rangle = \begin{pmatrix}1\\0\\0\end{pmatrix}, \quad \overrightarrow{Betty} = |e_2\rangle = \begin{pmatrix}0\\1\\0\end{pmatrix}, \quad \overrightarrow{Chris} = |e_3\rangle = \begin{pmatrix}0\\0\\1\end{pmatrix}.$$

Let the object vector space be $\mathbb{R}^n$ for some arbitrary $n \in \mathbb{N}$, where we take $|n_1\rangle$ to be the standard basis vector for $\mathbb{R}^n$ with 1 in the first position and 0 elsewhere. Let $\overrightarrow{holidays} = |n_1\rangle$. We will treat the concept *students* as being a hypernym of the individual students in our universe. In other words,

$$\ulcorner students \urcorner = \frac{1}{3}\ulcorner Annie \urcorner + \frac{1}{3}\ulcorner Betty \urcorner + \frac{1}{3}\ulcorner Chris \urcorner = \frac{1}{3}|e_1\rangle\langle e_1| + \frac{1}{3}|e_2\rangle\langle e_2| + \frac{1}{3}|e_3\rangle\langle e_3|.$$

Finally, let the verb $\overline{enjoy} \in \mathbb{R}^3 \otimes S \otimes \mathbb{R}^n$ be given by $\overline{enjoy} = \sum_{(i,j)\in R}|e_i\rangle|n_j\rangle$, where

$$R = R_{\text{enjoy}} = \{(i,j) |\ |e_i\rangle \text{ enjoys } |n_j\rangle\},$$

in the style of [7]. The corresponding density matrix for the verb is then:

$$\ulcorner enjoy \urcorner = \sum_{\substack{(i,j)\in R \\ (r,s)\in R}} |e_i\rangle\langle e_r| \otimes |n_j\rangle\langle n_s|.$$

Suppose that Annie and Betty are known to enjoy holidays, while Chris does not. Then the above set becomes simply $R = \{(1,1),(2,1)\}$.

Clearly, we have that $\ulcorner Annie \urcorner \preccurlyeq_{p_{max}} \ulcorner student \urcorner$ for $p_{max} = \frac{1}{3}$, since this is the maximum value of p for which we have $\ulcorner student \urcorner - p\ulcorner Ann \urcorner \succeq 0$, where:

$$\ulcorner student \urcorner - \frac{1}{3}\ulcorner Annie \urcorner = \begin{pmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} - \frac{1}{3}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix}.$$

We will see that the p-hyponymy for $p = \frac{1}{3}$ does translate into p-hyponymy of sentence $\ulcorner A \urcorner$ to sentence $\ulcorner B \urcorner$. However, we will also observe that this will no longer be the maximum value of p for which we have sentence hyponymy. First of all, consider the meanings of the two sentences:

$$\begin{aligned} \ulcorner A \urcorner &= (\varepsilon_N \otimes 1_S \otimes \varepsilon_N)(\ulcorner Annie \urcorner \otimes \ulcorner enjoy \urcorner \otimes \ulcorner holidays \urcorner) \\ &= (\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\left(|e_1\rangle\langle e_1| \otimes \sum_{\substack{(i,j)\in R \\ (r,s)\in R}} |e_i\rangle\langle e_r| \otimes |n_j\rangle\langle n_s| \otimes |n_1\rangle\langle n_1| \right) \\ &= \sum_{\substack{(i,j)\in R \\ (r,s)\in R}} \langle e_1|e_i\rangle\langle e_1|e_r\rangle\langle n_1|n_j\rangle\langle n_1|n_s\rangle = \sum_{\substack{(i,j)\in R \\ (r,s)\in R}} \delta_{1i}\,\delta_{1r}\,\delta_{1j}\,\delta_{1s} = \sum_{\substack{(1,1)\in R \\ (1,1)\in R}} 1 \quad = 1 \end{aligned} \tag{5.3}$$

$$\begin{aligned} \ulcorner B \urcorner &= (\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\,(\ulcorner student \urcorner \otimes \ulcorner enjoy \urcorner \otimes \ulcorner holidays \urcorner) \\ &= (\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\left(\frac{1}{3}\,(|e_1\rangle\langle e_1| + |e_2\rangle\langle e_2| + |e_3\rangle\langle e_3|) \otimes \sum_{\substack{(i,j)\in R \\ (r,s)\in R}} |e_i\rangle\langle e_r| \otimes |n_j\rangle\langle n_s| \otimes |n_1\rangle\langle n_1| \right) \\ &= \frac{1}{3}\left(\sum_{\substack{(i,1)\in R \\ (r,1)\in R}} \langle e_1|e_i\rangle\langle e_1|e_r\rangle + \sum_{\substack{(i,1)\in R \\ (r,1)\in R}} \langle e_2|e_i\rangle\langle e_2|e_r\rangle + \sum_{\substack{(i,1)\in R \\ (r,1)\in R}} \langle e_3|e_i\rangle\langle e_3|e_r\rangle \right) \\ &= \frac{1}{3}\left(\sum_{\substack{(i,1)\in R \\ (r,1)\in R}} \delta_{1i}\delta_{1r} + \sum_{\substack{(i,1)\in R \\ (r,1)\in R}} \delta_{2i}\delta_{2r} + \sum_{\substack{(i,1)\in R \\ (r,1)\in R}} \delta_{3i}\delta_{3r} \right) \\ &= \frac{1}{3}\left(\sum_{(1,1)\in R} 1 + \sum_{(2,1)\in R} 1 + 0 \right) = \frac{1}{3} \times 2 = \frac{2}{3} \end{aligned} \tag{5.4}$$

Clearly, we have that $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner$ for $p = \frac{1}{3}$, as $\frac{2}{3} - \frac{1}{3} \times 1 \geq 0$, but this is not the maximum value of p for which this p-hyponymy holds. The max value for which this works is $p = \frac{2}{3}$. The reason why this happens is that when we work with a truth-theoretic sentence space, the meanings of the sentences that we obtain in the end are trivial density matrices, i.e just one-dimensional, and hence they do not capture all the information that a non-trivial matrix can. In a sense, instead of obtaining a real density matrix meaning of the sentences, we just get the traces of density matrices.

Example 2: Simple case of object hyponymy

We now give a simple case with a non-truth-theoretic sentence space, in which we show that the p-hyponymy of the objects of two sentences translates into p-hyponymy between the sentences, and that the maximality of the value of p is also preserved.

Let $m \in \mathbb{N}$, $m > 2$ be such that $\{n_i\}_{i=1}^{m}$ is a collection of standard basis vectors for $\mathbb{R}^m$. We will use the nouns:

$$\overrightarrow{Gretel} = |n_1\rangle, \quad \overrightarrow{gingerbread} = |n_2\rangle, \quad \overrightarrow{cake} = |n_3\rangle, \quad \overrightarrow{pancakes} = |n_4\rangle,$$

with corresponding pure density matrices:

$$\ulcorner Gretel \urcorner = |n_1\rangle\langle n_1|, \quad \ulcorner gingerbread \urcorner = |n_2\rangle\langle n_2|, \quad \ulcorner cake \urcorner = |n_3\rangle\langle n_3|, \quad \ulcorner pancakes \urcorner = |n_4\rangle\langle n_4|$$

Let the mixed density matrix corresponding to the hypernym *sweets* be given by:

$$\ulcorner sweets \urcorner = \frac{1}{10}|n_2\rangle\langle n_2| + \sum_{i=3}^{m} p_i \, |n_i\rangle\langle n_i|.$$

Our object and subject vector space will be $\mathbb{R}^m$ and for the sentence space we take $S = \mathbb{R}^m \otimes \mathbb{R}^m$. We take the verb *like* to be given by $\overline{like} \in \mathbb{R}^m \otimes S \otimes \mathbb{R}^m$,

$$\overline{like} = \sum_{jk} C_{jk}^{like} |n_j\rangle|n_j\rangle|n_k\rangle|n_k\rangle,$$

where the coefficients C_{jk} give us the weight with which $|n_j\rangle$ likes $|n_k\rangle$.
For the rest of this example, we will adopt the following abuse of notation for the purpose of brevity:

$$|s_{jk}\rangle = |n_j\rangle|n_k\rangle, \quad \langle s_{jk}| = \langle n_j|\langle n_k|, \quad |s_{ij}\rangle\langle s_{kl}| = |n_i\rangle\langle n_k| \otimes |n_j\rangle\langle n_l|.$$

Then the density matrix representation of our verb becomes:

$$\begin{aligned} \ulcorner like \urcorner &= \left(\sum_{jk} C_{jk}^{like} \, |n_j\rangle|s_{jk}\rangle|n_k\rangle \right) \left(\sum_{lp} C_{lp}^{like} \, \langle n_l|\langle s_{lp}|\langle n_p| \right) \\ &= \sum_{jklp} C_{jk}^{like} C_{lp}^{like} |n_j\rangle\langle n_l| \otimes |s_{jk}\rangle\langle s_{lp}| \otimes |n_k\rangle\langle n_p|. \end{aligned} \tag{5.5}$$

We will consider the following two sentences:

A := Gretel likes sweets.
B := Gretel likes gingerbread.

Let the corresponding (density matrix) sentence meanings be given by:

$$\begin{aligned} \ulcorner A \urcorner &= (\varepsilon_N \otimes 1_S \otimes \varepsilon_N) \left(\ulcorner Gretel \urcorner \otimes \ulcorner like \urcorner \otimes \ulcorner sweets \urcorner \right) \\ \ulcorner B \urcorner &= (\varepsilon_N \otimes 1_S \otimes \varepsilon_N) \left(\ulcorner Gretel \urcorner \otimes like \urcorner \otimes \ulcorner gingerbread \urcorner \right) \end{aligned} \tag{5.6}$$

Observe that

$$\ulcorner gingerbread \urcorner \preccurlyeq_p \ulcorner sweets \urcorner \quad \text{for} \quad p \leq \frac{1}{10}.$$

In particular, we have p_{max}-hyponymy between *gingerbread* and *sweets* for $p_{max} = \frac{1}{10}$. We will now show that this hyponymy translates to the sentence level, as by **Corollary 2**. With $\varphi = \varepsilon_N \otimes 1_S \otimes \varepsilon_N$, we have,

$$\begin{aligned}
\ulcorner A \urcorner &= \varphi \left(|n_1\rangle\langle n_1| \otimes \sum_{jklp} C^{like}_{jk} C^{like}_{lp} \, |n_j\rangle\langle n_l| \otimes |s_{jk}\rangle\langle s_{lp}| \otimes |n_k\rangle\langle n_p| \otimes \left(\frac{1}{10} \, |n_2\rangle\langle n_2| + \sum_{i=3}^{m} p_i |n_i\rangle\langle n_i| \right) \right) \\
&= \frac{1}{10} \, \varphi \left(|n_1\rangle\langle n_1| \otimes \sum_{jklp} C^{like}_{jk} C^{like}_{lp} \, |n_j\rangle\langle n_l| \otimes |s_{jk}\rangle\langle s_{lp}| \otimes |n_p\rangle\langle n_p| \otimes |n_2\rangle\langle n_2| \right) + \\
&+ \varphi \left(|n_1\rangle\langle n_1| \otimes \sum_{jklp} C^{like}_{jk} C^{like}_{lp} \, |n_j\rangle\langle n_l| \otimes |s_{jk}\rangle\langle s_{lp}| \otimes |n_k\rangle\langle n_p| \otimes \sum_{i=3}^{m} p_i \, |n_i\rangle\langle n_i| \right) \\
\ulcorner B \urcorner &= \varphi \left(|n_1\rangle\langle n_1| \otimes \sum_{jklp} C^{like}_{jk} C^{like}_{lp} \, |n_j\rangle\langle n_l| \otimes |s_{jk}\rangle\langle s_{lp}| \otimes |n_k\rangle\langle n_p| \otimes |n_2\rangle\langle n_2| \right)
\end{aligned} \tag{5.7}$$

We claim that the maximum p-hyponymy between $\ulcorner B \urcorner$ and $\ulcorner A \urcorner$ is achieved for $p = \frac{1}{10}$. In other words, this is the maximum value of p for which we have $\ulcorner B \urcorner \preccurlyeq_p \ulcorner A \urcorner$, i.e. $\ulcorner A \urcorner - p \ulcorner B \urcorner \succeq 0$.
To see this, consider $\ulcorner A \urcorner - \frac{1}{10} \ulcorner B \urcorner$. We get:

$$\begin{aligned}
(\varepsilon_N \otimes 1_S \otimes \varepsilon_N) & \left(|n_1\rangle\langle n_1| \otimes \sum_{jklp} C^{like}_{jk} C^{like}_{lp} \, |n_j\rangle\langle n_l| \otimes |s_{jk}\rangle\langle s_{lp}| \otimes |n_k\rangle\langle n_p| \otimes \sum_{i=3}^{m} p_i \, |n_i\rangle\langle n_i| \right) \\
&= \sum_{ijklp} C^{like}_{jk} C^{like}_{lp} p_i \, \langle n_1|n_j\rangle\langle n_1|n_l\rangle\langle n_k|n_i\rangle\langle n_p|n_i\rangle \, |s_{jk}\rangle\langle s_{lp}| \\
&= \sum_{ijklp} C^{like}_{jk} C^{like}_{lp} p_i \, \delta_{1j}\delta_{1l}\delta_{ik}\delta_{ip} \, |s_{jk}\rangle\langle s_{lp}| \\
&= \sum_{i=3}^{m} C^{like}_{1i} C^{like}_{1i} p_i \, |s_{1i}\rangle\langle s_{1i}| \\
&= \sum_{i=3}^{m} \left(C^{like}_{1i} \right)^2 p_i \, |n_1\rangle\langle n_1| \otimes |n_i\rangle\langle n_i|
\end{aligned} \tag{5.8}$$

Now suppose that Gretel is the only object in our universe that likes sweets and that she only likes pancakes, cakes and gingerbread, all with equal weights, which we take to be $\frac{1}{3}$, i.e. we let:
$C^{like}_{jk} = \begin{cases} \frac{1}{3} & \text{if } j = 1, k \in \{2, 3, 4\} \\ 0 & \text{o.w.} \end{cases}$
Then (5.8) above becomes simply $\frac{1}{9} \left(p_3 |n_1\rangle\langle n_1| \otimes |n_3\rangle\langle n_3| + p_4 |n_1\rangle\langle n_1| \otimes |n_4\rangle\langle n_4| \right)$. Note that this is isomorphic to $\frac{1}{9} \left(p_3 |n_3\rangle\langle n_3| + p_4 |n_4\rangle\langle n_4| \right)$, and that the former is a positive semi-definite matrix if and only if the latter is. Alternatively, to simplify this example we could have just taken the object vector space to be one-dimensional and consisting only of Gretel, in which case $\overrightarrow{Gretel} = \overrightarrow{1}$ and we get the same outcome.

Since $|n_3\rangle\langle n_3|$ and $|n_4\rangle\langle n_4|$ are pure state and p_3 and p_4 are non-negative real numbers, we get that the matrix $p_3 |n_3\rangle\langle n_3| + p_4 |n_4\rangle\langle n_4|$ is positive semi-definite, which is what we claimed.

It is easy to see that for any other value of p below $\frac{1}{10}$, we would have also obtained a positive semi-definite matrix upon computing $\ulcorner A \urcorner - p \ulcorner B \urcorner$.

Example 4

Now suppose that the subject and object vector spaces are two-dimensional and spanned by $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. For convenience, we denote these vectors by $|e_1\rangle$ and $|e_2\rangle$ respectively when dealing with the subject vector space and $|n_1\rangle$ and $|n_2\rangle$ in the context of the object vector space. We let:

$$|e_1\rangle = \overrightarrow{Hansel}, \qquad |e_2\rangle = \overrightarrow{Gretel}, \qquad |n_1\rangle = \overrightarrow{gingerbread}, \qquad |n_2\rangle = \overrightarrow{cake}.$$

The density matrices for the hypernyms *the siblings* and *sweets* are:

$$\ulcorner \textit{the siblings} \urcorner = \frac{1}{2}\ulcorner \textit{Hansel} \urcorner + \frac{1}{2}\ulcorner \textit{Gretel} \urcorner \quad \text{and} \quad \ulcorner \textit{sweets} \urcorner = \frac{1}{2}\ulcorner \textit{gingerbread} \urcorner + \frac{1}{2}\ulcorner \textit{cake} \urcorner.$$

The verb *like* is given as before:

$$\ulcorner \textit{like} = \urcorner \sum_{ijkl} C_{ij}^{like} C_{kl}^{like} \, |e_i\rangle\langle e_k| \otimes |s_{ij}\rangle\langle s_{kl}| \otimes |n_j\rangle\langle n_l|,$$

where $C_{ij}^{like} = \begin{cases} 1 & \text{if} \;\; |e_i\rangle \; \textit{likes} \; |n_j\rangle \\ 0 & \text{o.w.} \end{cases}$
and we assume that Gretel likes gingerbread but not cake and Hansel likes both.

Then we have:

$$\begin{aligned}
\ulcorner A \urcorner &= (\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\left(\ulcorner \textit{Gretel} \urcorner \otimes \ulcorner \textit{like} \urcorner \otimes \ulcorner \textit{gingerbread} \urcorner\right) \\
\ulcorner B \urcorner &= (\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\left(\ulcorner \textit{the siblings} \urcorner \otimes \ulcorner \textit{like} \urcorner \otimes \ulcorner \textit{sweets} \urcorner\right) \\
&= (\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\left(\left(\frac{1}{2}\ulcorner \textit{Gretel} \urcorner + \frac{1}{2}\ulcorner \textit{Hansel} \urcorner\right) \otimes \ulcorner \textit{like} \urcorner \otimes \left(\frac{1}{2}\ulcorner \textit{gingerbread} \urcorner + \frac{1}{2}\ulcorner \textit{cake} \urcorner\right)\right) \\
&= \frac{1}{4}(\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\left(\ulcorner \textit{Gretel} \urcorner \otimes \ulcorner \textit{like} \urcorner \otimes \ulcorner \textit{gingerbread} \urcorner\right) + \\
&+ \frac{1}{4}(\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\left(\ulcorner \textit{Gretel} \urcorner \otimes \ulcorner \textit{like} \urcorner \otimes \ulcorner \textit{cake} \urcorner\right) + \\
&+ \frac{1}{4}(\varepsilon_N \otimes 1_S \otimes \varepsilon_N)\left(\ulcorner \textit{Hansel} \urcorner \otimes \ulcorner \textit{like} \urcorner \otimes \left(\ulcorner \textit{gingerbread} \urcorner + \ulcorner \textit{cake} \urcorner\right)\right)
\end{aligned} \tag{5.9}$$

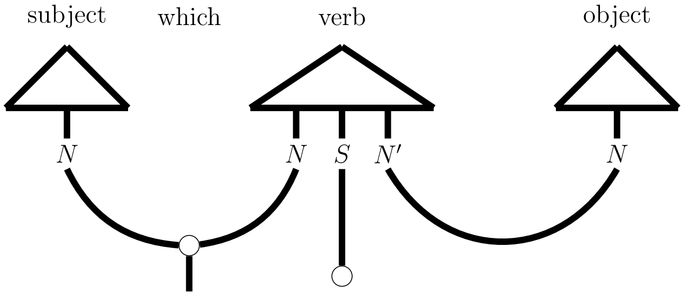
Clearly, $\ulcorner B \urcorner - \frac{1}{4}\ulcorner A \urcorner$ gives us just the last two lines of the above expression, which we compute explicitly to be:

$$\begin{aligned}
&\frac{1}{4}|s_{22}\rangle\langle s_{22}| + \frac{1}{4}|s_{11}\rangle\langle s_{11}| + \frac{1}{4}|s_{12}\rangle\langle s_{12}| \\
&\quad = \frac{1}{4}|n_2\rangle\langle n_2| \otimes |n_2\rangle\langle n_2| + \frac{1}{4}|n_1\rangle\langle n_1| \otimes |n_1\rangle\langle n_1| + \frac{1}{4}|n_1\rangle\langle n_1| \otimes |n_2\rangle\langle n_2| \\
&\quad = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{pmatrix} + \begin{pmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{pmatrix},
\end{aligned} \tag{5.10}$$

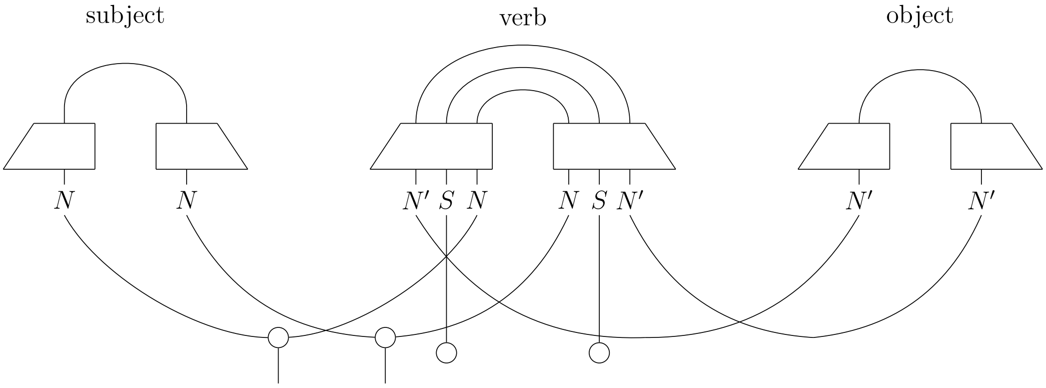
which is obviously a positive semi-definite matrix.

5.4.2 P-Hyponymy in relative clauses

The diagrammatic representation of subject relative clauses in CPM(**FHilb**) is:



which in **FHilb** looks like:



Without loss of generality assume that the relative pronoun is *which*. Then the meaning map for the relative clause *subject which verb object* in CPM(**FHilb**) is $\mu_N \otimes \iota_S \otimes \varepsilon_N$ and the meaning of the relative clause is given by:

$$(\mu_N \otimes \iota_S \otimes \varepsilon_N)(\ulcorner subj \urcorner \otimes \ulcorner verb \urcorner \otimes \ulcorner obj \urcorner).$$

Relationship between relative clauses '*A which verb C*' and '*B which verb D*' where $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner$ and $\ulcorner C \urcorner \preccurlyeq_q \ulcorner D \urcorner$

We obtain a result very similar to the one we had for the positive semi-definite sentence types, under the same assumptions.

Theorem 2. *Let A, B, C and D be nouns with corresponding density matrix representations $\ulcorner A \urcorner$, $\ulcorner B \urcorner$, $\ulcorner C \urcorner$ and $\ulcorner D \urcorner$, and such that $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner$ and $\ulcorner C \urcorner \preccurlyeq_q \ulcorner D \urcorner$, where $\ulcorner B \urcorner = p \ulcorner A \urcorner + \sum_i p_i \ulcorner X_i \urcorner$ and $\ulcorner D \urcorner = q \ulcorner C \urcorner + \sum_j q_j \ulcorner Y_j \urcorner$ for some $p, q \in (0, 1]$. Then we have that:*

$$\varphi\left(A\, which\, verb\, C\right) \preccurlyeq_{pq} \varphi\left(B\, which\, verb\, D\right).$$

Proof. The proof of this result is identical to that of the positive transitive sentence case, except for the fact that when we consider

$$\varphi(\ulcorner B \urcorner\, which\, verb\, \ulcorner D \urcorner) - pq\, \varphi(\ulcorner A \urcorner\, which\, verb\, \ulcorner C \urcorner)$$

we get $\varphi = \mu_N \otimes \iota_S \otimes \varepsilon_N$ applied to

$$p \ulcorner A \urcorner \otimes \ulcorner Z \urcorner \otimes \sum_j q_j \ulcorner Z_j \urcorner + \sum_i p_i \ulcorner X_i \urcorner \otimes \ulcorner Z \urcorner \otimes (q \ulcorner C \urcorner + \sum_j q_j \ulcorner Y_j \urcorner),$$

instead of $\varphi = (\varepsilon_N \otimes 1_S \otimes \varepsilon_N)$ applied to the same. The result is, however, still a positive quantity by the property of the morphisms μ_N, 1_S and ε_N to map density matrices to density matrices. Thus, we can conclude as before that:

$$\varphi\left(B\, which\, verb\, D\right) \preccurlyeq_{pq} \varphi\left(A\, which\, verb\, C\right).$$

□

Like before, we have the two special cases where $A = C$ or $B = D$, which are (respectively) the following two corollaries.

Corollary 3. *Let A, B, C we nouns with corresponding density matrix representations $\ulcorner A \urcorner$, $\ulcorner B \urcorner$ and $\ulcorner C \urcorner$ and such that $\ulcorner B \urcorner = p \ulcorner A \urcorner + \sum_i p_i \ulcorner X_i \urcorner$. Then we have that:*

$$\varphi(\ulcorner A \urcorner\, which\, verb\, \ulcorner C \urcorner) \preccurlyeq_p \varphi(\ulcorner B \urcorner\, which\, verb\, \ulcorner C \urcorner),$$

where $\varphi = \mu_N \otimes \iota_S \otimes \varepsilon_N$.

Corollary 4. *Let A, C, D be nouns with corresponding density matrix representations $\ulcorner A \urcorner$, $\ulcorner C \urcorner$ and $\ulcorner D \urcorner$ and such that $\ulcorner D \urcorner = p\ulcorner C \urcorner + \sum_j q_j \ulcorner Y_j \urcorner$. Then we have that:*

$$\varphi(\ulcorner A \urcorner \, which \, verb \, \ulcorner C \urcorner) \preccurlyeq_p \varphi(\ulcorner A \urcorner \, which \, verb \, \ulcorner D \urcorner) :$$

Then just like with positive transitive sentences, we can generalise these results to the cases where we just know that $\ulcorner A \urcorner \preccurlyeq_p \ulcorner B \urcorner$ and $\ulcorner C \urcorner \preccurlyeq_q \ulcorner D \urcorner$.

Given that this kind of transition from the word to the sentence level works in a very similar fashion here to the way in which it did in the positive transitive sentences gives an indication of the possibility for a more general result that applies to a broader class of sentences and other structures. We will come back to this observation in the next section. First we give an example with a relative clause.

Example

We will consider the containment of the sentence

A := Elderly ladies who own cats.

in the sentence

B := Women who own animals.

First of all, let the subject and object space for the vectors corresponding to the subjects and object of our sentences be $\mathbb{R}^2$ and $\mathbb{R}^3$ respectively. Let:

$$\overrightarrow{elderly\ ladies} = |e_1\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \overrightarrow{young\ ladies} = |e_2\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

and the density matrix for the hypernym *women* be:

$$\ulcorner women \urcorner = \frac{1}{3} \ulcorner elderly\ ladies \urcorner + \frac{2}{3} \ulcorner young\ ladies \urcorner = \frac{1}{3} |e_1\rangle\langle e_1| + \frac{2}{3} |e_2\rangle\langle e_2|.$$

Similarly, let:

$$\overrightarrow{cats} = |n_1\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \overrightarrow{dogs} = |n_2\rangle = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \overrightarrow{hamsters} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

and take the density matrix for *animals* to be:

$$\ulcorner animal \urcorner = \frac{1}{2}\ulcorner cats \urcorner + \frac{1}{4}\ulcorner dogs \urcorner + \frac{1}{4}\ulcorner hamsters \urcorner = \frac{1}{2} |n_1\rangle\langle n_1| + \frac{1}{4} |n_2\rangle\langle n_2| + \frac{1}{4} |n_3\rangle\langle n_3|$$

The sentence space will not matter in this case, as it gets deleted by the ι_S morphism, so we just take it to be an unspecified S. Let the verb *own* be given by $\overline{own} \in \mathbb{R}^2 \otimes S \otimes \mathbb{R}^3$,

$$\overline{own} = \sum_{ij} C_{ij} |e_i\rangle |s\rangle |n_j\rangle,$$

with corresponding density matrix

$$\ulcorner own \urcorner = \sum_{ijkl} C_{ij} C_{kl} \, |e_i\rangle\langle e_k| \otimes |s\rangle\langle s'| \otimes |n_j\rangle\langle n_l|.$$

Then the meaning of sentences A and B are given by:

$$\begin{aligned}
\ulcorner A \urcorner &= (\mu_N \otimes \iota_S \otimes \varepsilon_N)(\ulcorner elderly\ ladies \urcorner \otimes \ulcorner own \urcorner \otimes \ulcorner cats \urcorner) = (\mu_N \otimes \iota_S \otimes \varepsilon_N)(|e_1\rangle\langle e_1| \otimes \ulcorner own \urcorner \otimes |n_1\rangle\langle n_1|) \\
\ulcorner B \urcorner &= (\mu_N \otimes \iota_S \otimes \varepsilon_N)(\ulcorner women \urcorner \otimes \ulcorner own \urcorner \otimes \ulcorner animals \urcorner) \\
&= (\mu_N \otimes \iota_S \otimes \varepsilon_N) \left(\left(\frac{1}{3} |e_1\rangle\langle e_1| + \frac{2}{3} |e_2\rangle\langle e_2| \right) \otimes \ulcorner own \urcorner \otimes \left(\frac{1}{2} |n_1\rangle\langle n_1| + \frac{1}{4} |n_2\rangle\langle n_2| + \frac{1}{4} |n_3\rangle\langle n_3| \right) \right) \\
&= \frac{1}{6} (\mu_N \otimes \iota_S \otimes \varepsilon_N)(|e_1\rangle\langle e_1| \otimes \ulcorner own \urcorner \otimes |n_1\rangle\langle n_1|) + \\
&+ \frac{1}{12} (\mu_N \otimes \iota_S \otimes \varepsilon_N) (|e_1\rangle\langle e_1| \otimes \ulcorner own \urcorner \otimes (|n_2\rangle\langle n_2| + |n_3\rangle\langle n_3|)) + \\
&+ \frac{1}{3} (\mu_N \otimes \iota_S \otimes \varepsilon_N) \left(|e_2\rangle\langle e_2| \otimes \ulcorner own \urcorner \otimes (|n_1\rangle\langle n_1| + \frac{1}{2} |n_2\rangle\langle n_2| + \frac{1}{2} |n_3\rangle\langle n_3|) \right)
\end{aligned} \tag{5.11}$$

Clearly, $\ulcorner B \urcorner - \frac{1}{6}\ulcorner A \urcorner$ gives us just:

$$
\begin{aligned}
&\frac{1}{6}\,(\mu_N \otimes \iota_S \otimes \varepsilon_N)(|e_1\rangle\langle e_1| \otimes \ulcorner own \urcorner \otimes |n_1\rangle\langle n_1|)\\
&+ \frac{1}{12}\,(\mu_N \otimes \iota_S \otimes \varepsilon_N)\,(|e_1\rangle\langle e_1| \otimes \ulcorner own \urcorner \otimes (|n_2\rangle\langle n_2| + |n_3\rangle\langle n_3|))\\
&+ \frac{1}{3}\,(\mu_N \otimes \iota_S \otimes \varepsilon_N)\left(|e_2\rangle\langle e_2| \otimes \ulcorner own \urcorner \otimes (|n_1\rangle\langle n_1| + \frac{1}{2}\,|n_2\rangle\langle n_2| + \frac{1}{2}\,|n_3\rangle\langle n_3|)\right)\\
&= \frac{1}{12}\left(\sum_{ijkl} C_{ij}C_{kl}\,\langle e_1|e_i\rangle\langle e_1|e_k\rangle\langle n_2|n_j\rangle\langle n_2|n_l\rangle + \sum_{ijkl} C_{ij}C_{kl}\,\langle e_1|e_i\rangle\langle e_1|e_k\rangle\langle n_3|n_j\rangle\langle n_3|n_l\rangle\right)|e_1\rangle\langle e_1|\\
&+ \frac{1}{3}\left(\sum_{ijkl} C_{ij}C_{kl}\,\langle e_2|e_i\rangle\langle e_2|e_k\rangle\langle n_1|n_j\rangle\langle n_1|n_l\rangle + \frac{1}{2}\sum_{ijkl} C_{ij}C_{kl}\,\langle e_2|e_i\rangle\langle e_2|e_k\rangle\langle n_2|n_j\rangle\langle n_2|n_l\rangle + \right.\\
&\qquad\qquad \left. + \frac{1}{2}\sum_{ijkl} C_{ij}C_{kl}\,\langle e_2|e_i\rangle\langle e_2|e_k\rangle\langle n_3|n_j\rangle\langle n_3|n_l\rangle\right)|e_2\rangle\langle e_2|\\
&= \frac{1}{12}\,\left(C_{12}^2 + C_{13}^2\right)|e_1\rangle\langle e_1| + \left(\frac{1}{3}\,C_{21}^2 + \frac{1}{6}\,C_{22}^2 + \frac{1}{6}\,C_{23}^2\right)|e_2\rangle\langle e_2|\\
&= \alpha\,|e_1\rangle\langle e_1| + \beta\,|e_2\rangle\langle e_2|
\end{aligned}
$$

Here $\alpha = \frac{1}{12}(C_{12}^2 + C_{13}^2)$ and $\beta = \frac{1}{3}\,C_{21}^2 + \frac{1}{6}\,C_{22}^2 + \frac{1}{6}\,C_{23}^2$ are both non-negative and so regardless of the actual values of the verb coefficients, we always get a linear combination of non-negative scalar multiples of two density matrices: $|e_1\rangle\langle e_1|$ and $|e_2\rangle\langle e_2|$. Thus, the resulting matrix is necessarily positive semi-definite. In other words,

$$\ulcorner B \urcorner - \frac{1}{6}\ulcorner A \urcorner \succeq 0.$$

5.4.3 General case of P-Hyponymy

As we already observed, it seems like there is no reason to believe that the p-hyponymy result lifted to the sentence level should not be more generally applicable to all sorts of sentence structures. The result below is meant to show that.

In the theorem below, we adopt the following conventions:

- A *positive sentence* is assumed to be a sentence that does not contain any negations, including words like *not* and nouns which are in some way the opposite of other nouns (in the case where these two appear in different sentences), such as *satisfaction* and *dissatisfaction* or antonyms.
- Adjective-noun pairs are counted as one word whose meaning is assumed to have been computed. Note that the output of the meaning map applied to an adjective-noun phrase is a noun type and hence we can safely make this assumption. Hence, for simplicity, adjective-noun pairs will be called nouns for the purposes of the result and proof below.
- The *sentence length* of a sentence or a noun phrase is the number of words in it, not counting definite and indefinite articles and assuming that a noun modified by an adjective is counted as one word.

Theorem 3 (Generalised Sentence P-Hyponymy). *Let Φ and Ψ be two positive sentences of the same sentence length and type, containing some or all of the following: nouns, verbs, relative pronouns (who/that/which/whom) and possessive pronouns (whose). Let S be the common sentence space for Φ and Ψ (if they contain any verbs), and assume that S is not truth-theoretic. Denote the nouns and verbs 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 $\ulcorner A_1 \urcorner, \ldots, \ulcorner A_n \urcorner$ and $\ulcorner B_1 \urcorner, \ldots, \ulcorner B_n \urcorner$ respectively. Suppose that $\ulcorner A_{i_1} \urcorner \preccurlyeq_{p_{i_1}} \ulcorner B_{i_1} \urcorner, \ldots, \ulcorner A_{i_l} \urcorner \preccurlyeq_{p_{i_l}} \ulcorner B_{i_l} \urcorner$ for some subset $\{i_1, \ldots, i_l\} \subseteq$*

$\{1,\ldots,n\}$ and some $p_{i_1},\ldots,p_{i_l} \in (0,1]$, and that $\ulcorner A_k \urcorner = \ulcorner B_k \urcorner$ for $k \in [1,n]$, $k \notin \{i_1,\ldots,i_l\}$. 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 we have that:

$$\varphi(\Phi) \preccurlyeq_{p_{i_i}\cdots p_{i_l}} \varphi(\Psi).$$

Intuitively, this means that if (some of) the functional words of a sentence Φ are p-hyponyms of (some of) the functional words of sentence Ψ, then this hyponymy is translated into sentence hyponymy. Moreover, the strength of the sentence hyponymy can only be as strong as the combined hyponymy of the individual words, where we express combined hyponymy as multiplication.

In real terms, we can think of this as follows. Suppose that we have two sentences which are identical apart from one word in each, say word A in the first sentence and word B in the second. Suppose, furthermore, that A is a p-max hyponym of B for some p. We can think of this p as the proportion of sentences that use word B which can be replaced by sentences which use word A. For example, *soap opera* should be a p-hyponym of *TV show* for some value of p. This value of p helps us determine how often the sentence '*High school students watch TV shows*' can be replaced by '*High school students watch soap operas*'. Now, clearly, the more hyponym-hypernym pairs we have in the two sentences, the more they differ from each other and hence the less likely it is that we can use one sentence instead of the other. This is captured in the fact that the strengths of the hyponymy of the individual hyponym-hypernym pairs multiply together to give us, essentially, the extent to which the sentence containing the hyponyms can replace the sentence containing the hypernyms.

Before proceeding with the proof, we also observe that because of our assumptions of the sentence structures that we allow, the meaning map φ can only be comprised of parallel and/or sequential morphisms from the following list: $\{\varepsilon,\eta,\nu,\iota,\mu,\zeta,\Delta,1\}$, which can be reduced to just $\{\varepsilon,\mu,\iota,\Delta,1\}$. These maps are sufficient for the purposes of modeling the meaning of sentences of the above types. It is possible that the meanings of other sentence types can also be expressed using the morphisms of CPM(**FHilb**), but since this constitutes work in progress, we do not consider these at the moment.

For the proof below we will assume w.l.o.g. that $i_j = j$, $\forall i_j \in \{i_1,\ldots,i_n\}$, so that $\ulcorner A_1 \urcorner \preccurlyeq_{p_1} \ulcorner B_1 \urcorner$, … , $\ulcorner A_l \urcorner \preccurlyeq_{p_l} \ulcorner B_l \urcorner$ for some $l \leq n$ and $\ulcorner A_{l+1} \urcorner = \ulcorner B_{l+1} \urcorner, \ldots, \ulcorner A_n \urcorner = \ulcorner B_n \urcorner$. In other words, we will assume that any hyponymy that occurs on the word level between words from the two sentences happens between consecutive words. In the case where we have hyponymy between non-consecutive words, the proof is similar, but slightly more notationally involved.

Proof. First of all, we have $\ulcorner A_i \urcorner \preccurlyeq_{p_i} \ulcorner B_i \urcorner$ for $i \in [1,l]$ for some $l \leq n$. This means that for each i, we have density matrices X_{i_j} and non-negative reals p_{i_j} such that $\ulcorner B_i \urcorner = p_i \ulcorner A_i \urcorner + \sum_j p_{i_j} \ulcorner X_j \urcorner$. Let $\ulcorner Y_i \urcorner = \sum_j p_{i_j} \ulcorner X_j \urcorner$.

Now consider the meanings of the two sentences. We have:

$$\begin{aligned}
\varphi(\Phi) &= \phi(\ulcorner A_1 \urcorner \otimes \ldots \otimes \ulcorner A_n \urcorner), \\
\varphi(\Psi) &= \varphi(\ulcorner B_1 \urcorner \otimes \ldots \otimes \ulcorner B_n \urcorner) \\
&= \varphi\left((p_1 \ulcorner A_1 + \ulcorner Y_1 \urcorner) \otimes \ldots \otimes (p_l \ulcorner A_l \urcorner + \ulcorner Y_l \urcorner) \otimes \ulcorner B_{l+1} \urcorner \otimes \ldots \otimes \ulcorner B_n \urcorner\right)
\end{aligned} \tag{5.12}$$

Before proceeding, we first establish some convenient notation. Let $i_1,\ldots,i_l \in \{0,1\}$ be binary values and let $A_{i_1\cdots i_l}$ be the string of tensors of $\ulcorner A_p \urcorner$'s and $\ulcorner Y_q \urcorner$'s such that for each $k,m \in [1,l]$, $\ulcorner A_k \urcorner$ is in the string and in position number k iff $i_k = 1$ and $\ulcorner Y_m \urcorner$ is in the string and in position number m iff $i_m = 0$. For example if $l = 4$, then $A_{1010} = \ulcorner A_1 \urcorner \otimes \ulcorner Y_2 \urcorner \otimes \ulcorner A_3 \urcorner \otimes \ulcorner Y_4 \urcorner$ and $A_{0001} = \ulcorner Y_1 \urcorner \otimes \ulcorner Y_2 \urcorner \otimes \ulcorner Y_3 \urcorner \otimes \ulcorner A_4 \urcorner$. We also set $A_{1\ldots1} = \ulcorner A_1 \urcorner \otimes \ldots \ulcorner A_l \urcorner = 0$ for any value of l. Similarly, we let $P_{i_1\cdots i_l}$ to be the string of p_m's such that for each $k \in [1,l]$, p_k is in the string iff $i_k = 1$ and the number 1 is in the string iff $i_k = 0$. Thus, $P_{1010} = p_1 \times 1 \times p_3 \times 1 = p_1 p_3$ and $P_{0001} = 1 \times 1 \times 1 \times 1 \times p_4 = p_4$.

With this notation, (5.12) becomes:

$$\begin{aligned}
\varphi\left(\Psi\right) = \varphi &\left(\sum_{i_1,\ldots,i_l \in \{0,1\}} P_{i_1,\ldots,i_l} A_{i_1 \ldots i_l} \otimes \ulcorner B_{l+1} \urcorner \otimes \cdots \otimes \ulcorner B_n \urcorner \right) \\
&\qquad\qquad + \varphi(p_1 \ulcorner A_1 \urcorner \otimes \ldots \otimes p_l \ulcorner A_l \urcorner \otimes \ulcorner B_{l+1} \urcorner \otimes \ldots \otimes \ulcorner B_n \urcorner) \\
= &\sum_{i_1,\ldots,i_l \in \{0,1\}} P_{i_1,\ldots,i_l}\, \varphi\left(A_{i_1,\ldots,i_l} \otimes \ulcorner A_{l+1} \urcorner \otimes \ldots \otimes \ulcorner A_n \urcorner\right) \\
&\qquad\qquad + p_1 \cdots p_l\, \varphi(\ulcorner A_1 \urcorner \otimes \ldots \otimes \ulcorner A_l \urcorner \otimes \ulcorner A_{l+1} \urcorner \otimes \ldots \otimes A_n \urcorner)
\end{aligned}$$

Finally, consider $\varphi(\Psi) - p_1 \cdots p_l\, \varphi(\Phi)$, for which we get:

$$\sum_{i_1,\ldots,i_l \in \{0,1\}} P_{i_1,\ldots,i_l}\, \varphi\left(A_{i_1,\ldots,i_l} \otimes \ulcorner A_{l+1} \urcorner \otimes \ldots \otimes \ulcorner A_n \urcorner\right). \tag{5.13}$$

Now, since all of the matrices $\ulcorner A_i \urcorner$ and $\ulcorner X_{i_j} \urcorner$ are density matrices by assumption, and since φ is a completely positive map, we get that each $\varphi\left(A_{i_1 \ldots i_l} \otimes \ulcorner A_{l+1} \urcorner \otimes \ldots \otimes \ulcorner A_n \urcorner\right)$ is a positive semi-definite matrix. All the p_k's are non-negative and hence so is any product of any subcollection of these. Thus, (5.13) is a sum of non-negative scalar multiples of positive semi-definite matrices, and as such is itself a positive semi-definite matrix. In other words, $(\varphi(\Psi) - p_1 \cdots p_l \varphi(\Phi)) \geq 0$. We conclude that:

$$\varphi(\Phi) \preccurlyeq_{p_1 \cdots p_l} \varphi(\Psi),$$

as claimed. □

5.5 Using the same idea for other applications

The idea behind the mathematical structure of p-hyponymy can be potentially applied to other scenarios where we are not necessarily interested in determining to what extent one concept is contained in another. For example, we may have two concepts which are not in a hypernym-hyponym relationship but are still similar to each other in some kind of asymmetric way. Consider again the China - North Korea example from the end of the previous chapter. China and North Korea are by no means synonyms and neither of them is a hyponym of the other. However, as we noted earlier, North Korea does get judged as being more similar to China than vice versa. We want to somehow represent this without having to use the verb *is similar to.*

The following is one possible way in which asymmetrical similarity can be captured in the case where we have concepts that can be represented with respect to the same set of salient features.

Let $\ulcorner \Phi \urcorner$ and $\ulcorner \Psi \urcorner$ be two mixed states in CPM(**FHilb**) represented by:

$$\ulcorner \Phi \urcorner = \sum_{i=1}^{n} \alpha_i \ulcorner A_i \urcorner \quad \text{and} \quad \ulcorner \Psi \urcorner = \sum_{j=1}^{n} \beta_j \ulcorner A_j \urcorner,$$

where $\{\ulcorner A_i \urcorner\}_{i=1}^{n}$ is a set of pure states representing salient features which form an orthonormal basis for the space of $n \times n$ real square matrices. Also, $\alpha_i, \beta_j \in (0, 1]$. We do not impose here the normalising condition that $\sum_i \alpha_i = \sum_i \beta_i = 1$.

Define the **degree of similarity of $\ulcorner \Phi \urcorner$ with respect to $\ulcorner \Psi \urcorner$** to be the average of the maximum values of p_i for which we have:

$$\ulcorner \alpha_i\, A_i \urcorner \preccurlyeq_{p_i} \ulcorner \beta_i\, A_i \urcorner, \tag{5.14}$$

defined as before to mean $\ulcorner \beta_i\, A_i \urcorner - p_i \ulcorner \alpha_i\, A_i \urcorner \succeq 0$. Note that we require that each $p_i \geq 0$ but we no longer impose an upper bound on the value of p_i. More formally, we write:

$$\mathscr{S}\left(\ulcorner \Phi \urcorner, \ulcorner \Psi \urcorner\right) = avg\left(\{p_i \,|\, \ulcorner \alpha_i\, A_i \urcorner \preccurlyeq_{p_i} \ulcorner \beta_i\, A_i \urcorner \, \text{and} \; \forall i, \nexists q_i \; : \; \ulcorner \alpha_i\, A_i \urcorner \preccurlyeq_{q_i} \ulcorner \beta_i\, A_i \urcorner \text{ and } q_i > p_i\}\right) \tag{5.15}$$

Clearly, the maximum value for which (5.14) holds is $p_i = \frac{\beta_i}{\alpha_i}$, since

$$\ulcorner \beta_i\, A_i \urcorner - p \ulcorner \alpha_i\, A_i \urcorner \succeq 0 \iff (\beta_i - p\, \alpha_i) \ulcorner A_i \urcorner \succeq 0 \iff \beta_i - p\, \alpha_i \geq 0 \iff p \leq \frac{\beta_i}{\alpha_i}.$$

So we can alternatively write (5.15) as

$$\mathscr{S}\left(\ulcorner \Phi \urcorner, \ulcorner \Psi \urcorner\right) = avg\left(\frac{\beta_1}{\alpha_1}, \ldots, \frac{\beta_n}{\alpha_n}\right).$$

We see that for each i, the number p_i is a ratio of the presence of feature A_i in concept Ψ to the presence of the same feature in concept Φ. We could think of these as some kind of 'prominence ratio values'. Then the effect of averaging over all of these is to obtain an average ratio of the features of one concept to the other. Performing this calculation in both directions allows us to compare the two concepts based on which of them has, on average, more prominent features than the other one.

To see how this works in practice, consider again the example of China and North Korea. Take the set of the basic (salient) features to be the same as before and let:

$$|e_1\rangle = \overrightarrow{big}, \quad |e_2\rangle = \overrightarrow{populous}, \quad |e_3\rangle = \overrightarrow{prominent}, \quad |e_4\rangle = \overrightarrow{affluent}$$
$$|e_5\rangle = \overrightarrow{East\ Asian}, \quad |e_6\rangle = \overrightarrow{communist}, \quad |e_7\rangle = \overrightarrow{militarised},$$

where $\{|e_i\rangle\}_{i=1}^{7}$ is the standard orthonormal basis for $\mathbb{R}^7$. Let $\ulcorner E_i \urcorner = |e_1\rangle\langle e_i|$ for each $i \in [1, 7]$ be the pure state in CPM(**FHilb**) corresponding to the basis vectors. Represent the two countries as:

$$\ulcorner China \urcorner = 0.9 \ulcorner E_1 \urcorner + 1 \ulcorner E_2 \urcorner + 0.8 \ulcorner E_3 \urcorner + 0.5 \ulcorner E_4 \urcorner + 1 \ulcorner E_5 \urcorner + 0.6 \ulcorner E_6 \urcorner + 0.6 \ulcorner E_7 \urcorner$$
$$\ulcorner NorthKorea \urcorner = 0.3 \ulcorner E_1 \urcorner + 0.4 \ulcorner E_2 \urcorner + 0.4 \ulcorner E_3 \urcorner + 0.2 \ulcorner E_4 \urcorner + 0.1 \ulcorner E_5 \urcorner + 0.8 \ulcorner E_6 \urcorner + 0.9 \ulcorner E_7 \urcorner$$

The relative degrees of similarity are then given by:

$$\mathscr{C}\left(\ulcorner China \urcorner, \ulcorner North\ Korea \urcorner\right) = avg\left(\frac{0.3}{0.9}, \frac{0.4}{1}, \frac{0.4}{0.8}, \frac{0.2}{0.5}, \frac{0.1}{1}, \frac{0.8}{0.6}, \frac{0.9}{0.6}\right) \approx 0.88$$
$$\mathscr{C}\left(\ulcorner North\ Korea \urcorner, \ulcorner China \urcorner\right) = avg\left(\frac{0.9}{0.3}, \frac{1}{0.4}, \frac{0.8}{0.4}, \frac{0.5}{0.2}, \frac{1}{0.1}, \frac{0.6}{0.8}, \frac{0.6}{0.9}\right) \approx 3.1$$

Thus we see that the degree of similarity between North Korea and China is greater that the degree of similarity between China and North Korea. This confirms the result that we had before, and in fact gives an even stronger indication of asymmetry, based on the relative prominence of the features of the two countries.

This is a very simple idea and we do not need density matrices to implement it. However, the advantage of working with density matrices is that we can now do the same as with p-hyponymy and translate this asymmetric relationship into the same one but between sentences containing these words, i.e one sentence containing the word China and one containing the words North Korea which are of the same type. This can further be extended to cover sentences where apart from the words China and North Korea we also have other words in a hyponym-hypernym relationship.

Chapter 6

Conclusion and future directions

In this dissertation we make use of the framework of the distributional compositional model of meaning to capture examples of the asymmetry that naturally occurs in phenomena from linguistics and cognition.

Part 1: Overextension with respect to concept combination and the pet fish phenomenon

We first built upon work done in [25] in order to consider an alternative representation of overextension with respect to concept combination, exemplified by the Pet Fish phenomenon, in a distributional compositional model of meaning that uses density matrices as containers of word meanings in CPM(**FHilb**) instead of vectors in **FHilb**. This approach has the advantage of allowing us to see more clearly the correlation and interaction between the features of the parent concept and its parts. The density matrix for pet fish is then compared to the pure matrix for goldfish via the fidelity measure to produce a more intuitive output for similarity than the one obtained when comparing the vectors of these concepts via the cosine measure.

Work on the Pet Fish problem, both in [25] and here, is so far purely theoretical and based on the assumption that the hand-chosen set of salient features with respect to which we model the concepts is suitable. Empirical evidence will be necessary to establish the appropriateness of the vector-based and the density matrix-based models and determine if there is any significant advantage of the latter over the former, given that the complexity of the computation doubles with the passage from vectors to matrices. We would also need to test it on other examples, such as the case where we have a concept which is comprised of two completely unrelated to each other words, like *school furniture*.

Part 2: Asymmetry of similarity judgment via positive transitive sentences containing the verb *is similar to*

We briefly considered another cognitive phenomenon that goes by the name of asymmetry of similarity judgment. We showed how the simple original framework of [7] for modeling the meaning of positive transitive sentences can be used to account for the fact that the sentence '*North Korea is similar to China*' is judged by human subjects to be more likely than '*China is similar to North Korea*' [40]. This was done by taking a graded truth-theoretic sentence space and representing the asymmetric verb *is similar to* with respect to the same features as those used for the the construction of the country vectors. The meanings of the two sentences are then simply real numbers and we judge the higher of these to correspond to the more likely sentence. Again, for a true indication of the validity of this model, some experimental support will be needed. It will also be useful to normalise the values obtained in the computation of the meanings so that outputs of the meaning maps are between 0 and 1, thus allowing us to better judge the degree to which one concept is more similar to the other one than vice versa.

Part 3: P-Hyponymy

In the final chapter of this dissertation, we presented a new way of measuring the relative hyponymy of one concept to another. We called this measure p-hyponymy and defined it for values of p in the range $(0, 1]$. We think of these p-values as being essentially probabilities of being able to replace the concept A with B in some context. We were primarily interested in the upper bound on the p-values which we called p-max hyponymy. We showed how this measure gives an order on the density matrices that are used to represent hypernyms and how this can be lifted to the sentence level.

In our theoretical model, we assumed that a hypernym B can be expressed as a weighted mixture of the density matrices corresponding to a relevant set of its hyponyms, where the weights can be extracted from a large body of text and based on contextual co-occurrence. However, it might be the case that this representation of the weights does not give us a reliable indication of the strength of the hyponymy and we need a better method for constructing them. Also, a large-scale experiment would be needed to establish the connection between having p-hyponymy between A and B, and any implications this carries about the entailment of A in B. We worked under the assumption that all the matrices we have are normalised - proper density matrices. We should consider what happens if we drop this assumption. In particular, if we allow values for the weights of the individual words that make up a density matrix of a hypernym to exceed 1, and we don't normalise, we might end up with p-hyponymy for p greater than 1. Whether or not this is at all indicative of anything interesting from a linguistic point of view is currently unclear.

We established the result that the p-values of the p-hyponymy between hyponym-hypernym pairs in two sentences A and B multiply together to give us the combined p-hyponymy between A and B. We interpreted this as being the probability of being able to replace sentence B by sentence A in some larger context. Intuitively, the more hyponyms we have in A with corresponding hypernyms in B, the narrower the scope for replacing B with A in some larger body of text. However, assuming that this result applies in the same way to all grammatical structures A and B seems like an unnatural over-simplification. For example, suppose that the p-max hyponym between the words *cat* and *animal* with respect to some context is 0.5. Then this is lifted to p-hyponymy for $p = 0.5$ between the sentences '*Cats like milk*' and '*Animals likes milk*' and between '*cats which are fluffy*' and '*animals that are fluffy*'. One possible way to model such differences in sentence structure lies in making use of the non-maximal values of hyponymy in some way, ideally dependent on the context of the sentences and other relevant factors. How this can be done in practice is a possible avenue for future work.

We stated at the start of Chapter 5 that it might be useful to think of hyponymy as being a weaker version of typicality, or even prototypicality. One way of applying this idea in practice could be as follows. Starting from the knowledge that concept A is a p-hyponym of concept B for some value of p, we first determine the largest such value in the $(0, 1]$ range, by computing eigenvalues or otherwise. We can then experimentally establish threshold values for hyponymy and typicality, say ε and ζ such that if $\varepsilon \leq p \leq \zeta$ then we conclude that we have hyponymy, while if $p > \zeta$, we have typicality. These thresholds can be set by examining the results obtained for an appropriate number of hyponym-hypernym pairs and comparing these against data on the relative hyponymy or typicality between these pairs extracted via other means, such as experiments carried out with target groups.

We concluded with a possible modification to the p-hypomymy measure that allowed us to consider the China - North Korea example from Chapter 4 in a way that does not require the use of the verb *is similar to* at all. We leave the further development and possible applications of this modified measure to future work.

References

[1] Samson Abramsky and Nikos Tzevelekos. Introduction to categories and categorical logic. In *New structures for physics*, pages 3–94. Springer, 2011.

[2] Esma Balkır. Using density matrices in a compositional distributional model of meaning. 2014.

[3] Michael Barr and Charles Wells. *Category theory for computing science*, volume 49. Prentice Hall New York, 1990.

[4] Stephen Clark and Stephen Pulman. Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium: Quantum Interaction*, pages 52–55, 2007.

[5] Bob Coecke and Aleks Kissinger. Picturing quantum processes. In *Book of Abstracts*, page 4, 2014.

[6] Bob Coecke and Éric Oliver Paquette. Categories for the practising physicist. In *New Structures for Physics*, pages 173–286. Springer, 2011.

[7] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*, 2010.

[8] J. R. Firth. A synopsis of linguistic theory 1930-55. 1952-59:1–32, 1957.

[9] J.R. Firth. *A Synopsis of Linguistic Theory, 1930-1955.* 1957.

[10] Rebecca Green, Carol A Bean, and Sung Hyon Myaeng. *The semantics of relationships: an interdisciplinary perspective*, volume 3. Springer Science & Business Media, 2013.

[11] Edward Grefenstette. Category-theoretic quantitative compositional distributional models of natural language semantics. *arXiv preprint arXiv:1311.1539*, 2013.

[12] Edward Grefenstette and Mehrnoosh Sadrzadeh. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404. Association for Computational Linguistics, 2011.

[13] Edward Grefenstette and Mehrnoosh Sadrzadeh. Experimenting with transitive verbs in a discocat. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 62–66. Association for Computational Linguistics, 2011.

[14] Edward Grefenstette, Mehrnoosh Sadrzadeh, Stephen Clark, Bob Coecke, and Stephen Pulman. Concrete sentence spaces for compositional distributional models of meaning. In *Computing Meaning*, pages 71–86. Springer, 2014.

[15] James A Hampton. Inheritance of attributes in natural concept conjunctions. *Memory & Cognition*, 15(1):55–71, 1987.

[16] James A Hampton. Disjunction of natural concepts. *Memory & Cognition*, 16(6):579–591, 1988.

[17] James A Hampton. Overextension of conjunctive concepts: Evidence for a unitary model of concept typicality and class inclusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1):12, 1988.

[18] Chris Heunen and Jamie Vicary. Lectures on categorical quantum mechanics. *Computer Science Department. Oxford University*, 2012.

[19] Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, Stephen Pulman, and Bob Coecke. Reasoning about meaning in natural language with compact closed categories and frobenius algebras. *CoRR*, abs/1401.5980, 2014.

[20] Joachim Kock. *Frobenius algebras and 2-d topological quantum field theories*, volume 59. Cambridge University Press, 2004.

[21] J Lambek and C Casadio. Computational algebraic approaches to natural language. *Polimetrica, Milan*, 2006.

[22] Joachim Lambek. The mathematics of sentence structure. *American mathematical monthly*, pages 154–170, 1958.

[23] Joachim Lambek. Type grammar revisited. In *Logical aspects of computational linguistics*, pages 1–27. Springer, 1999.

[24] Joachim Lambek. From word to sentence. *Polimetrica, Milan*, 2008.

[25] Martha Lewis and Bob Coecke. A compositional explanation of the 'pet fish' phenomenon. 2015.

[26] Saunders Mac Lane. *Categories for the working mathematician*, volume 5. Springer Science & Business Media, 1978.

[27] Richard Montague. English as a formal language. 1970.

[28] Richard Montague. Universal grammar. *1974*, pages 222–46, 1970.

[29] Daniel N Osherson and Edward E Smith. On the adequacy of prototype theory as a theory of concepts. *Cognition*, 9(1):35–58, 1981.

[30] Barbara Partee. Compositionality. *Varieties of formal semantics*, 3:281–311, 1984.

[31] Roger Penrose. Applications of negative dimensional tensors. *Combinatorial mathematics and its applications*, 221244, 1971.

[32] Robin Piedeleu. Ambiguity in categorical models of meaning. 2014.

[33] Robin Piedeleu, Dimitri Kartsaklis, Bob Coecke, and Mehrnoosh Sadrzadeh. Open system categorical quantum semantics in natural language processing. *arXiv preprint arXiv:1502.00831*, 2015.

[34] Mehrnoosh Sadrzadeh, Stephen Clark, and Bob Coecke. The frobenius anatomy of word meanings i: subject and object relative pronouns. *Journal of Logic and Computation*, 23(6):1293–1317, 2013.

[35] Mehrnoosh Sadrzadeh, Stephen Clark, and Bob Coecke. The frobenius anatomy of word meanings ii: possessive relative pronouns. *Journal of Logic and Computation*, page exu027, 2014.

[36] Mehrnoosh Sadrzadeh and Edward Grefenstette. A compositional distributional semantics, two concrete constructions, and some experimental evaluations. *CoRR*, abs/1105.1702, 2011.

[37] Peter Selinger. Dagger compact closed categories and completely positive maps. *Electronic Notes in Theoretical Computer Science*, 170:139–163, 2007.

[38] Peter Selinger. A survey of graphical languages for monoidal categories. In *New structures for physics*, pages 289–355. Springer, 2011.

[39] Paul Smolensky and Géraldine Legendre. The harmonic mind: From neural computation to optimality-theoretic grammar (vol. 1: Cognitive architecture). 2006.

[40] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.