
A Quantum-Inspired Analysis of Human Disambiguation Processes

Foundational Theory and Applications

Daphne Pauline Wang

Thesis submitted as part of:
PhD in Computer Science



University College London

Declaration

I, Daphne Pauline Wang, confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Formal languages are essential for computer programming and are constructed to be easily processed by computers. In contrast, natural languages are much more challenging and instigated the field of Natural Language Processing (NLP). One major obstacle is the ubiquity of ambiguities. Recent advances in NLP have led to the development of large language models, which can resolve ambiguities with high accuracy. At the same time, quantum computers have gained much attention in recent years as they can solve some computational problems faster than classical computers. This new computing paradigm has reached the fields of machine learning and NLP, where hybrid classical-quantum learning algorithms have emerged. However, more research is needed to identify which NLP tasks could benefit from a genuine quantum advantage. In this thesis, we applied formalisms arising from foundational quantum mechanics, such as contextuality and causality, to study ambiguities arising from linguistics. By doing so, we also reproduced psycholinguistic results relating to the human disambiguation process. These results were subsequently used to predict human behaviour and outperformed current NLP methods.

Impact Statement

Large Language Models (LLMs) are used in everyday life nowadays, especially with applications such as ChatGPT. These LLMs are data-hungry, which creates reproducibility and environmental issues. Furthermore, they act as a “black box”, and whether they learn the same language features as humans is unclear.

In recent years, research in quantum information theory has led to machine learning algorithms using near-term quantum computers. These advances offer the possibility of using quantum computers for NLP purposes. However, results about quantum advantages in these near-term learning algorithms are yet to be found.

This project aimed to address some of these issues. The impact of this work is mainly academic, where our contributions spanned multiple fields, including Natural Language Processing, Quantum Computing, and Linguistics. The project also has the potential to develop a new generation of more cognitively plausible learning algorithms.

Human and quantum processes

This project aimed to bring forward similarities between linguistic and physical phenomena using a shared mathematical language. Drawing a parallel between linguistic and quantum mechanical concepts may help identify which NLP tasks could benefit from quantum resources. The non-determinism of natural language ambiguities notably offered a promising place to start.

Regarding cognitive plausibility issues, we adopted a foundational approach and looked at human disambiguating strategies. The framework identified essential structures of the disambiguation process, aligning with psycholinguistic theories.

We then used the established structure to simulate some disambiguation processes using near-term quantum computers.

An interdisciplinary project

This project involved elements from various research fields, including psycholinguistics, artificial intelligence, and quantum computing, and opened new avenues for research in each of them.

Natural Language Processing This project provides an alternative, more cognitively plausible, and transparent approach to NLP. By exploiting the structure of the human disambiguation process and the computational power of quantum systems, we produced a learning algorithm that delivers results even with a small training dataset.

Quantum Computing In this work, we provide a strategy for finding applications of quantum computing in areas that are, on the surface, unrelated to quantum mechanics. For instance, this project provided the first instance of quantum(-like) contextuality in linguistic data, where contextuality is a fundamental distinction between classical and quantum statistics.

Linguistics By introducing the mathematical tools from quantum physics, we provide new tools to study linguistic phenomena and cognitive processes. Using these mathematical frameworks, we can create new models of cognitive processes that are, by design, easy to simulate using quantum resources.

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. For a research manuscript that has already been published (if not yet published, please skip to section 2):

(a) **What is the title of the manuscript?**

On the Quantum-like Contextuality of Ambiguous Phrases

(b) **Please include a link to or doi for the work:**

<https://aclanthology.org/2021.semSPACE-1.5.pdf>

(c) **Where was the work published?**

ACL Anthology

(d) **Who published the work?**

Association for Computational Linguistics

(e) **When was the work published?**

June 2021

(f) **List the manuscript's authors in the order they appear on the publication:**

Daphne Wang, Mehrnoosh Sadrzadeh, Samson Abramsky, Víctor Cervantes

(g) **Was the work peer reviewed?**

Yes

(h) **Have you retained the copyright?**

Yes

(i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

Yes

<https://arxiv.org/abs/2107.14589>

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3):

(a) **What is the current title of the manuscript?**

(b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**

If 'Yes', please please give a link or doi:

(c) **Where is the work intended to be published?**

(d) **List the manuscript's authors in the intended authorship order:**

(e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4):

- **Daphne Wang:** Development of the mathematical framework; Data collection; Data analysis and computations.
- **Mehrnoosh Sadrzadeh:** Computational linguistic expertise; Data collection; Supervision.
- **Samson Abramsky:** Contribution of the original idea; sheaf-theoretic contextuality expertise.
- **Víctor Cervantes:** Data analysis and computations; Contextuality-by-Default expertise.

4. **In which chapter(s) of your thesis can this material be found?**

Chapter 3

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):

Candidate:

Date: 05/12/2023

Supervisor/Senior Author signature (where appropriate):

Date: 05/12/2023

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. For a research manuscript that has already been published (if not yet published, please skip to section 2):

(a) **What is the title of the manuscript?**

Analysing Ambiguous Nouns and Verbs with Quantum Contextuality Tools

(b) **Please include a link to or doi for the work:**

<https://doi.org/10.17791/jcs.2021.22.3.391>

(c) **Where was the work published?**

Journal of Cognitive Science

(d) **Who published the work?**

Journal of Cognitive Science

(e) **When was the work published?**

July 2021

(f) **List the manuscript's authors in the order they appear on the publication:**

Daphne Wang, Mehrnoosh Sadrzadeh, Samson Abramsky, Víctor H Cervantes

(g) **Was the work peer reviewed?**

Yes

(h) **Have you retained the copyright?**

Yes

(i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

No

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3):

- (a) **What is the current title of the manuscript?**
- (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**
If 'Yes', please please give a link or doi:
- (c) **Where is the work intended to be published?**
- (d) **List the manuscript's authors in the intended authorship order:**
- (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4):

- **Daphne Wang:** Development of the mathematical framework; Data collection; Data analysis and computations.
- **Mehrnoosh Sadrzadeh:** Computational linguistic expertise; Data collection; Supervision.
- **Samson Abramsky:** Contribution of the original idea; sheaf-theoretic contextuality expertise.
- **Víctor H. Cervantes:** Data analysis and computations; Contextuality-by-Default expertise.

4. **In which chapter(s) of your thesis can this material be found?**

Chapter 3

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):

Candidate

Date: 05/12/2023

Supervisor/Senior Author signature (where appropriate):

Date: 05/12/2023

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2):

(a) **What is the title of the manuscript?**

The Causal Structure of Semantic Ambiguities

(b) **Please include a link to or doi for the work:**

<https://doi.org/10.4204/EPTCS.394.12>

(c) **Where was the work published?**

Electronic Proceeding in Theoretical Computer Science

(d) **Who published the work?**

Electronic Proceeding in Theoretical Computer Science

(e) **When was the work published?**

November 2023

(f) **List the manuscript's authors in the order they appear on the publication:**

Daphne Wang, Mehrnoosh Sadrzadeh

(g) **Was the work peer reviewed?**

Yes

(h) **Have you retained the copyright?**

Yes

(i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

Yes

<https://arxiv.org/abs/2206.06807>

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3):

- (a) **What is the current title of the manuscript?**
- (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?**

If 'Yes', please please give a link or doi:

- (c) **Where is the work intended to be published?**
- (d) **List the manuscript's authors in the intended authorship order:**
- (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4):

- **Daphne Wang:** Development of the mathematical framework; Data analysis and computations.
- **Mehrnoosh Sadrzadeh:** Computational linguistic expertise; Supervision.

4. **In which chapter(s) of your thesis can this material be found?**

Chapter 3

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):

Candidate:

Date: 05/12/2023

Supervisor/Senior Author signature (where appropriate):

Date: 05/12/2023

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. For a research manuscript that has already been published (if not yet published, please skip to section 2):

(a) **What is the title of the manuscript?**

Causality and Signalling of Garden-Path Sentences

(b) **Please include a link to or doi for the work:**

<https://doi.org/10.1098/rsta.2023.0013>

(c) **Where was the work published?**

Philosophical Transactions of Royal Society A

(d) **Who published the work?**

Royal Society

(e) **When was the work published?**

January 2024

(f) **List the manuscript's authors in the order they appear on the publication:**

Daphne Wang, Mehrnoosh Sadrzadeh

(g) **Was the work peer reviewed?**

Yes

(h) **Have you retained the copyright?**

Yes

(i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**

No

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3):

- (a) **What is the current title of the manuscript?**
- (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'? If 'Yes', please give a link or doi:**
- (c) **Where is the work intended to be published?**
- (d) **List the manuscript's authors in the intended authorship order:**
- (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4):

- **Daphne Wang:** Development of the mathematical framework; Data analysis and computations.
- **Mehrnoosh Sadrzadeh:** Computational linguistic expertise; Supervision.

4. **In which chapter(s) of your thesis can this material be found?**

Chapter 5 & 6

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):

Candidate:

Date: 17/05/2024

Supervisor/Senior Author signature (where appropriate):

Date: 17/05/2024

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof Mehrnoosh Sadrzadeh, for her invaluable help throughout my PhD. I am incredibly lucky to benefit from knowledge and experience, and she has always encouraged me to achieve more. I could not have wished to have a better supervisor. I would also like to thank my examiners Fabio Zanasi and Rui Soares-Barbosa for their very helpful comments on the thesis. In addition, I would like to give special thanks to the many people who have also guided me through this journey, including Samson Abramsky, Ehtibar Dzhafarov, Shane Mansfield, Ruth Kempson, Wing Yee Chow, and Richard Breheny. I am also grateful to my office mates, Lachlan McPheat, Kin Ian Lo, and Hadi Wazni, for the many discussions we have had (and the countless hours of procrastination).

I also could not have done it without my husband, Joel Ingram, who kept me sane throughout the years. He has been my constant source of comfort and happiness. Additionally, I would like to thank my parents for their incredible influence on my life and emotional support. Finally, I would like to mention my friends and family from home, who have always believed in me more than I believed in myself.

To Popo.

CONTENTS

Introduction	1
I Background	7
1 Quantum theory and applications	9
1.1 A crash course in Category Theory	9
1.1.1 Basics of Category Theory	10
1.1.2 Sheaf theory	15
1.1.3 Monoidal categories	18
1.2 Describing Quantum Correlations	26
1.2.1 Contextuality	27
1.2.2 The sheaf-theoretic view of contextuality	36
1.2.3 The Contextuality-by-Default framework	46
1.3 Quantum Mechanics as a Process Theory	55
1.3.1 Features of quantum processes	55
1.3.2 Mixed states and density matrices	58
1.3.3 Causality in quantum processes	62
2 Ambiguities in Natural Languages	65
2.1 Lexical ambiguity in linguistics	66
2.1.1 The challenge of word-sense disambiguation	67
2.1.2 The human disambiguation process	76
2.2 Human parsing & garden-path sentences	80
2.2.1 Psycholinguistics theories of parsing	81

2.2.2	Surprisal predictions for garden-path sentences	85
II	Lexical Ambiguity	89
3	Aspects of the lexical disambiguation process	91
3.1	Methodology	92
3.1.1	The corpus dataset	94
3.1.2	The human judgment dataset	96
3.2	On the quantum-like contextuality of ambiguous phrases	99
3.2.1	Cyclic models of rank 2	100
3.2.2	Cyclic models of rank 4	104
3.3	Degrees of signalling and the levels of ambiguity	108
3.3.1	Cyclic systems of rank 2	109
3.3.2	Cyclic systems of rank 4	114
3.3.3	Discussion of the results	117
3.4	The causality of the disambiguation process	118
3.4.1	The direction of signalling in SV and VO models	120
3.4.2	Models with different levels of ambiguity	121
4	Quantum simulations of the disambiguation process	127
4.1	Methodology	128
4.1.1	The ansatz	128
4.1.2	The training process	130
4.1.3	Convergence	131
4.2	The prediction power of the variational circuits	134
4.2.1	Methods	134
4.2.2	Results	135
4.3	Obtaining quantum word embeddings	137
4.3.1	Methods	137
4.3.2	Results	140
4.4	Entanglement of phrases and words	141
4.4.1	Entanglement measures	141
4.4.2	Entanglement of the optimised circuits	142
4.4.3	Entanglement of the noun-embeddings	143

<i>CONTENTS</i>	xix
III Syntactic ambiguity	147
5 Modeling the human parsing process	149
5.1 A sheaf-theoretic model of the syntax of sentences	150
5.1.1 Incrementality	151
5.1.2 Grammatical structures	152
5.1.3 Probability distributions	156
5.2 Contextuality, causality and signalling of the models	158
5.2.1 Contextuality	158
5.2.2 Causality and signalling	158
5.2.3 Computing SF	159
6 Predicting garden-path effects	163
6.1 Methods	163
6.1.1 Tools	164
6.1.2 Method	165
6.1.3 Description of the datasets	167
6.2 Analysis of the predictions	171
6.2.1 The Sturt et al. dataset	172
6.2.2 The Grodner et al. dataset	177
6.2.3 General discussion	182
6.3 Comparison with surprisal	184
Conclusion	189
Bibliography	194
IV Appendix	215
A Sections of a presheaf and the sheafification of a presheaf	217
B Proof of the CHSH inequality	221
C Original proofs	225
C.1 Proof of proposition 1.44	225
C.2 Proof of proposition 1.46	227

C.3	Equivalence between causality notions	235
C.4	Proof of proposition 3.12	238
C.5	Proof of proposition 4.4	240
C.6	Proof of proposition 5.10	245
D	Lexical ambiguity dataset	249
D.1	List of ambiguous words	249
D.2	The corpus dataset	262
D.3	The human judgment dataset	283
D.4	Prediction dataset	290

INTRODUCTION

In this research program, we investigate some properties of the English language using mathematical tools from Quantum Mechanics. We created quantum-inspired models of human disambiguation processes from linguistic data. Using these models, we provide promising evidence that this method leads to novel quantum computing methods for Natural Language Processing.

Computational Linguistics

Artificial Intelligence (AI), or how to perform intelligent tasks done by humans algorithmically, is a longstanding challenge in Computer Science. Amongst one of the most widely studied areas in AI is the field of Natural Language Processing (NLP), whose goal is to understand natural language. This field is currently widely dominated by the use of Large Language Models (LLMs), such as BERT or GPT, which consist of artificial neural networks trained over large corpora. These LLMs are incredibly successful in various NLP tasks, such as text generation or knowledge extraction. There are, however, several criticisms of them, including:

- **Lack of reproducibility**, due to the immense resources needed for training;
- **Transparency**, i.e. the impossibility of tracing back the decision process of the neural network;
- **Cognitive plausibility**, i.e. whether these neural networks reproduce the way humans learn and process natural language data.

In parallel, computational linguistics aims to use computational tools to study human cognitive processes and natural languages. Computational linguistics and

NLP are highly related, and their distinction is increasingly blurred.

The aim of this thesis is more in line with computational linguistics since we investigate features of the natural language using tools from NLP (e.g. LLMs). The thesis focuses on English, although our approach could be replicated in other languages.

Quantum Computing

In 1994, Peter Shor published his famous article describing a quantum algorithm that can factorise an integer in polynomial time [166], thus demonstrating the use of quantum systems to solve hard computational tasks. The discovery of this algorithm sparked the interest of the computer science research community in quantum information theory and quantum computing. The idea behind quantum computations is straightforward. Instead of using bits – as in classical computing – information is encoded as *qubits*. Qubits can not only take values the values $|0\rangle$ or $|1\rangle$ but can also be expressed as the (complex) linear combination:

$$\alpha |0\rangle + \beta |1\rangle \quad \text{such that} \quad |\alpha|^2 + |\beta|^2 = 1 \quad (1)$$

Similarly, instead of considering operations between strings of bits as computations, we use operations between systems of qubits, which are subject to the laws of quantum mechanics.

Since Shor’s algorithms, it has been shown that quantum systems are capable of achieving speed-ups in various computational tasks in theory, in tasks such as database search [86], optimisation tasks [60] or simulation of physical systems [24, 16], and more recently in practice, in tasks such as sampling from a random quantum circuit [15] or boson sampling [202, 129]. In addition, although quantum advantages are more difficult to prove, quantum computations have started to find applications in various fields of computer science, including optimisation problems (e.g. [59, 60]) and AI (e.g. [65, 201, 182, 157]).

This project

Most research on quantum computing applications to AI and NLP consists of creating a quantum version of existing algorithms. Therefore, these approaches also suffer from the same issues as the classical approaches, including the need for more transparency and cognitive plausibility. In addition, the advantage of using quantum computing resources is not always clear and usually relies on heuristics.

In this work, we aim to address these problems for quantum NLP. By studying linguistic data using the formalisms of quantum mechanics, we create a parallel between linguistic phenomena and quantum systems, from which we can identify which features of natural languages would benefit from simulations on quantum hardware. We also show that, by using the mathematical frameworks developed to study quantum mechanics, we can uncover properties of the human disambiguation process.

This thesis uses the mathematics of *category theory*, in particular the notions of *sheaves* and *presheaves*. The main reason for doing so is the level of abstraction allowed by category theory, allowing us to draw parallels between quantum and linguistic systems. A second motivation is the *categorical quantum mechanics* research project which originated from the seminal paper of Samson Abramsky and Bob Coecke [7]. Indeed, the line of research showed that quantum mechanics can very elegantly be described in categorical terms (see more details in Chapter 1).

On top of that, this description has also been applied to various aspects of linguistics, notably in the Distributional Compositional Categorical models of meanings (also known as DisCoCat), which emanated from [42], or as semantics of Discourse Representation Theory [8].

This thesis is an additional example of the application of categorical quantum mechanics in linguistics to ambiguities in the English language. Ambiguities in English occur at different levels, from words to discourses. This project investigates the disambiguation process of two types of ambiguities, namely:

- **Lexical ambiguity** which happens when a single word has multiple interpretations. For example, the word *bank*, which could refer to a financial institution or the bank of a river;
- **Syntactic ambiguity** which happens when a phrase can have multiple grammatical structures. For example, in the sentence *She saw a man with binoculars*,

the phrase *with binoculars* can either attach to the verb *saw* or to the noun-phrase *a man*.

In the case of lexical ambiguity, we have studied the statistics of the meaning activations of subject-verb and verb-object phrases, where each word is lexically ambiguous. We then used the mathematical framework arising from the study of quantum contextuality and causality to study these statistics. We were able to show that the observed statistics from lexical ambiguity data show are capable to exhibit quantum-like contextuality. In addition, we were able to rederive some psycholinguistics results that were originally based on eye-tracking data (which is not easily reproducible and expensive to obtain). Using these results, we produced quantum simulations of the disambiguation process of subject-verb and verb-object phrases, which could then be applied to NLP tasks.

Regarding syntactic ambiguity, we also used the sheaf-theoretic frameworks originating from quantum contextuality and causality to create a model of the syntactic parsing process. This model was then used to make reading time predictions in special sentences, known as garden-path sentences. These predictions were closer to the human baseline than the ones obtained from state-of-the-art methods of computational linguistics.

Outline of the thesis

The aim of Part I is to introduce the concept we will use in the rest of the thesis.

- In Chapter 1, we introduce the different branches of categorical quantum mechanics that we will use in the subsequent parts.
- In Chapter 2, we introduce the main literature regarding lexical (Section 2.1) and syntactic (Section 2.2) ambiguities. In particular, we will describe the psycholinguistic theories relating to these ambiguity types and the computational tools that have been used in the past in NLP and computational linguistics.

Elements of Section 2.1 will be used in Part II, while theories introduced in Section 2.2 will be used in Part III.

Parts II and III correspond to the original contributions of the thesis. These parts can be read independently.

In Part II, we focus on the lexical disambiguation process.

- We start by studying the features of lexically ambiguous phrases in terms of quantum *contextuality* and *causality* in Chapter 3.
- In Chapter 4, we use the conclusions of Chapter 3 to generate a quantum model of the human disambiguation process using variational quantum circuits.

In Part III, we study the human parsing process by looking at *garden-path sentences*.

- Inspired by the psycholinguistic theories described in Section 2.2, we describe our sheaf-theoretic model of the human parsing process in Chapter 5.
- In Chapter 6, we evaluate the models from Chapter 5 empirically. We then compare the model's predictions with those from state-of-the-art computational linguistics models.

Published contributions

Several original contributions presented in thesis were published before the submission of this thesis. These are the following:

- Title: *On the Quantum-like Contextuality of Ambiguous Phrases*
Authors (publication order): **D. Wang**, M. Sadrzadeh, S. Abramsky and V. H. Cervantes
Published in: *ACL Anthology* as part of the *Proceedings of the 2021 Workshop on Semantic Spaces at the Intersection of NLP, Physics, and Cognitive Science (SemSpace)*
Publication date: 2021
Material presented in thesis in: Section 3.2
- Title: *Analysing Ambiguous Nouns and Verbs with Quantum Contextuality Tools*
Authors (publication order): **D. Wang**, M. Sadrzadeh, S. Abramsky and V. H. Cervantes
Published in: *Journal of Cognitive Science*
Publication date: 2021
Material presented in thesis in: Section 3.3

- Title: *Causality and Signalling of Garden-Path Sentences*
Authors (publication order): **D. Wang** and M. Sadrzadeh
Published in: *Philosophical Transaction of Royal Society A*
Publication date: 2024
Material presented in thesis in: Part III (Chapters 5 & 6)

Part I

BACKGROUND

Chapter 1

QUANTUM THEORY AND APPLICATIONS

This chapter introduces the quantum physics formalisms we will use in Parts II and III. In particular, we will make use of the categorical description of quantum mechanics. We will, therefore, start by introducing the mathematics of category theory in Section 1.1. In Section 1.2, we will describe quantum correlations, notably in terms of sheaf theory, and in Section 1.3, we will give a categorical description of quantum processes.

1.1 A crash course in Category Theory

Category theory originated in the work of Eilenberg and MacLane for homological algebra [57]. The field of category theory then rapidly evolved and reached other areas of mathematics such as algebraic geometry [85], set theory [115], as well as computer science [113, 66] and physics [7, 17].

In this section we introduce the basic concepts at the heart of category theory. This introduction is by no means comprehensive, and we will only present the elements that will be useful in subsequent chapters.

1.1.1 Basics of Category Theory

In this subsection, we start by defining the main notions of category theory, which will be subsequently expanded in Section 1.1.2 and Section 1.1.3.

Definition 1.1 (Category). A *category* \mathcal{C} consists of:

1. A collection of *objects* denoted as $ob(\mathcal{C})$
2. A set^[1] of *morphisms* for each pair of objects $A, B \in ob(\mathcal{C})$, denoted as $\mathcal{C}(A, B)$ equipped with:
 - (a) Sequential composition: for morphisms $f \in \mathcal{C}(A, B)$ and $g \in \mathcal{C}(B, C)$ we have $g \circ f \in \mathcal{C}(A, C)$.
 - (b) Identity morphism: for each object $A \in ob(\mathcal{C})$, there exists a unique morphism $id_A \in \mathcal{C}(A, A)$.

satisfying the following properties:

- (a) For any $f \in \mathcal{C}(A, B)$, $g \in \mathcal{C}(B, C)$ and $k \in \mathcal{C}(C, D)$:

$$k \circ (g \circ f) = (k \circ g) \circ f \quad (1.1)$$

In other words, sequential composition is associative.

- (b) For any $f \in \mathcal{C}(A, B)$ and $g \in \mathcal{C}(C, A)$:

$$f \circ id_A = f \quad id_A \circ g = g \quad (1.2)$$

Example 1.2. Here are standard examples of categories:

- a. The category of sets and functions denoted as **Sets**. The objects of **Sets** are sets, and morphisms are functions between sets. Composition is the standard functional composition, and identity morphisms are identity functions:

$$id_A \in \mathbf{Sets}(A, A) = a \mapsto a \quad (1.3)$$

^[1]We are here only considering locally small categories; in general categories, $\mathcal{C}(A, B)$ could be a proper class.

- b. The category of sets and relations denoted as \mathbf{Rel} . The objects are the same as the ones of \mathbf{Sets} , and the morphisms are binary relations, i.e. $\mathbf{Rel}(A, B) = \mathcal{P}(A \times B)$. The composition of relations $u \in \mathbf{Rel}(A, B)$ and $v \in \mathbf{Rel}(B, C)$ is defined as:

$$v \circ u \in \mathbf{Rel}(A, C) = \{(a, c) \mid \exists b \in B. (a, b) \in u \wedge (b, c) \in v\} \quad (1.4)$$

The identity morphisms in \mathbf{Rel} are then defined as:

$$id_A \in \mathbf{Rel}(A, A) = \{(a, a) \mid a \in A\} \quad (1.5)$$

- c. The category of vector spaces denoted as \mathbf{Vect} . Its objects are vector spaces, and morphisms between two vector spaces are linear maps. Composition is the standard composition of (linear) maps. We can also define the subcategory \mathbf{FdVect} of \mathbf{Vect} in which the objects are finite-dimensional vector spaces, and morphisms are defined as in \mathbf{Vect} . In \mathbf{FdVect} , morphisms can be seen as matrices (by fixing a basis), and composition as matrix multiplication. Similarly, the identity morphisms become identity matrices.
- d. A category is said to be a *preorder* if there is at most one morphism between two objects. This notion generalises the notion of order by taking the existence of a morphism between A and B to represent $A \leq B$. The presence of composition morphisms in a preorder means that the relation \leq is transitive, i.e. if $A \leq B$ and $B \leq C$ then $A \leq C$. Similarly, the existence of identity morphism indicates that the relation \leq is reflexive, i.e. $A \leq A$ for any object A .
- e. A preorder is said to be a *partial order* (or a *poset*) iff $A \leq B$ and $B \leq A$ implies that $A = B$. Furthermore, a partial order is a *total order* iff for any pair of distinct objects A and B , either $A \leq B$ or $B \leq A$.

It is often convenient to denote morphisms $f \in \mathcal{C}(A, B)$ as arrows:

$$A \xrightarrow{f} B \quad \text{or equivalently} \quad f : A \rightarrow B$$

Then, from associativity equation (1.1), the composition of several morphisms, say $f \in \mathcal{C}(A, B)$, $g \in \mathcal{C}(B, C)$, $h \in \mathcal{C}(C, D)$ and $k \in \mathcal{C}(D, E)$ can unambiguously be

written as:

$$A \xrightarrow{f} B \xrightarrow{g} C \xrightarrow{h} D \xrightarrow{k} E$$

In addition, identity morphisms can be omitted from equation (1.2).

From this, we can define the notion of a *diagram* in a category \mathcal{C} , which corresponds to a (labelled) directed graph such that nodes are objects in \mathcal{C} and (labelled) arrows $A \xrightarrow{f} B$ correspond to the morphism $f \in \mathcal{C}(A, B)$. Paths, therefore, correspond to the composition of morphisms. We then say that a diagram *commutes* iff the paths having the same endpoints are equal. For example, the commutativity of the following diagram represents the equation $g \circ f = l \circ k \circ h$:

$$\begin{array}{ccc} A & \xrightarrow{f} & B \\ h \downarrow & & \downarrow g \\ C & \xrightarrow{k} E \xrightarrow{l} & D \end{array}$$

In addition, it can be seen that given a set of arrows of a category, reversing all of the arrows still leads to a valid category. This category is known as the *opposite category*.

Definition 1.3. Given a category \mathcal{C} , its *opposite category*, denoted as \mathcal{C}^{op} , is defined as follows:

- The objects of \mathcal{C}^{op} are the same as the objects of \mathcal{C} ;
- Each morphism $f \in \mathcal{C}(A, B)$ gives a morphism $\tilde{f} \in \mathcal{C}^{op}(B, A)$. The composition $g \circ f \in \mathcal{C}(A, C)$, where $f \in \mathcal{C}(A, B)$ and $g \in \mathcal{C}(B, C)$, then gives $\widetilde{g \circ f} = \tilde{f} \circ \tilde{g} \in \mathcal{C}^{op}(C, A)$.

So far, given a category \mathcal{C} , the notion of “sameness” of objects is only captured as the equality of objects. However, equality might be too restrictive in general. For example, in **Sets**, the two sets $A_1 = \{0, 1, 2\}$ and $A_2 = \{1, 2, 3\}$ are different, but they have the same expressive power in the sense that any map $f_1 : A_1 \rightarrow B$ can be translated into a map $f_2 : A_2 \rightarrow B$ and conversely. In the category **Sets**, this notion is conveyed by the existence of a *bijection* between the sets A_1 and A_2 , and this notion extends to an arbitrary category as the notion of *isomorphism*.

Definition 1.4. A morphism $f : A \rightarrow B$ in a category \mathcal{C} is an *isomorphism* if there exists a morphism $f^{-1} : B \rightarrow A$ such that:

$$f \circ f^{-1} = id_B \quad \text{and} \quad f^{-1} \circ f = id_A \quad (1.6)$$

Example 1.5. a. As expected, the isomorphisms in **Sets** are the bijections.

b. In **Vect**, the isomorphisms are the isomorphisms of vector spaces.

c. In partial orders, the only isomorphisms are the identity morphisms.

Up to now, we have studied categories individually. We now look at relationships *between categories*.

Definition 1.6. A *functor* $\mathcal{F} : \mathcal{C} \rightarrow \mathcal{D}$ between two categories \mathcal{C} and \mathcal{D} is defined as follows:

- For each object A in the category \mathcal{C} gives an object $\mathcal{F}A$ in \mathcal{D} under the action of \mathcal{F} .
- Similarly, each morphism $f \in \mathcal{C}(A, B)$ gives a morphism $\mathcal{F}f \in \mathcal{D}(\mathcal{F}A, \mathcal{F}B)$ satisfying:

$$\mathcal{F}id_A = id_{\mathcal{F}A} \quad (1.7)$$

$$\mathcal{F}(g \circ f) = \mathcal{F}g \circ \mathcal{F}f \quad (1.8)$$

for every object A in \mathcal{C} , and any arrows f, g in \mathcal{C} .

Example 1.7. Let us look at some examples of functors.

- a. We can define a functor $F : \mathbf{Sets} \rightarrow \mathbf{Rel}$ which sends any set $A \in ob(\mathbf{Sets})$ to the same set $A \in ob(\mathbf{Rel})$. The action on morphisms in **Sets** (i.e. functions), will simply become the equivalent relation on **Rel**. Namely, for any $f : A \rightarrow B$ in **Sets**, we get:

$$Ff = \{(a, f(a)) \mid a \in A\} : A \rightarrow B$$

in **Rel**.

b. We can define a functor $\mathcal{D}_{\mathbb{R}_+} : \mathbf{Sets} \rightarrow \mathbf{Sets}$ which is defined on object as:

$$\mathcal{D}_{\mathbb{R}_+} :: U \mapsto \{d : U \rightarrow \mathbb{R}_+ \mid d \text{ is a probability distribution over } U\}$$

For any morphism $f : U \rightarrow V$, we then define $\mathcal{D}_{\mathbb{R}_+} f$ as:

$$\mathcal{D}_{\mathbb{R}_+} f :: d_U \mapsto d_V \text{ such that } d_V(v) = \sum_{u \in f^{-1}(v)} d_U(u)$$

The functor $\mathcal{D}_{\mathbb{R}_+}$ is called the *distribution monad*.

c. For any set A , we can define the *free monoid* over A , denoted as A^* which correspond to the set of lists of elements in A . In other words, $A^* \cong \coprod_{n \in \mathbb{N}} A^n$. This correspondence can be extended to a functor $F : \mathbf{Sets} \rightarrow \mathbf{Mon}$, where \mathbf{Mon} is the category of monoids and monoid homomorphisms. In this functor, any function $f : A \rightarrow B$ in \mathbf{Sets} is mapped to:

$$Ff : A^* \rightarrow B^* :: (a_1, a_2, \dots) \mapsto (f(a_1), f(a_2), \dots)$$

We can, moreover, study relationships between functors by looking at *natural transformations*.

Definition 1.8. Given two functor $\mathcal{F}, \mathcal{G} : \mathcal{C} \rightarrow \mathcal{D}$, a *natural transformation* $\eta : \mathcal{F} \Rightarrow \mathcal{G}$ is a family of maps $\{\eta_A : \mathcal{F}A \rightarrow \mathcal{G}A\}_{A \in \text{ob}(\mathcal{C})}$ such that for any morphism $f : A \rightarrow B$ in \mathcal{C} the following (naturality) square commutes:

$$\begin{array}{ccc} \mathcal{F}A & \xrightarrow{\mathcal{F}f} & \mathcal{F}B \\ \eta_A \downarrow & & \downarrow \eta_B \\ \mathcal{G}A & \xrightarrow{\mathcal{G}f} & \mathcal{G}B \end{array} \quad (1.9)$$

In addition, a natural transformation is a *natural isomorphism* if the morphisms η_A are isomorphisms for all $A \in \text{ob}(\mathcal{C})$.

Using the natural transformations as “morphisms” between functors also leads to the notion of functor category.

Definition 1.9 (Functor category). Given two categories \mathcal{C} and \mathcal{D} , we define the *functor category* $[\mathcal{C}, \mathcal{D}]$ (also written $\mathcal{D}^{\mathcal{C}}$) where:

- Objects of $[\mathcal{C}, \mathcal{D}]$ are functors $\mathcal{F} : \mathcal{C} \rightarrow \mathcal{D}$.
- Morphisms are natural transformations, and compositions are defined point-wise, i.e. $\mu \circ \eta = \{\mu_A \circ \eta_A \mid A \in \text{ob}(\mathcal{C})\}$.

1.1.2 Sheaf theory

In this project, we will make use of sheaves and presheaves. The general idea is that presheaves and sheaves define a mathematical notion of *consistency*. These concepts will be at the core of description of sheaf-theoretic contextuality described in Section 1.2.1, and subsequently used in the models developed in Part II and III. Here, we review the main definitions of sheaf theory.

Definition 1.10. Given a category \mathcal{C} , we define a *presheaf* over \mathcal{C} as a functor $\mathcal{F} : \mathcal{C}^{op} \rightarrow \mathbf{Set}$.

Remark 1.11. The above definition corresponds to the notion of set-valued presheaf. Depending on the school of thought, the term presheaf may also refer to the more general abelian presheaves or presheaves of modules, which take values in abelian groups or modules respectively instead of sets. Here, we will only consider set-valued sheaves and presheaves.

This work will mainly look at presheaves and sheaves over topological spaces. A topological space is usually defined as a tuple (X, τ) where X is a set of *points* and $\tau \subseteq \mathcal{P}(X)$ is the set of *open sets* which contains the empty set and is closed under arbitrary unions and finite intersections.

Remark 1.12. The class of topological spaces then extends to the category \mathbf{Top} where objects are topological spaces and morphisms are continuous functions between them.

Given a topological space $\mathcal{X} = (X, \tau)$, we can define a preorder category $\mathcal{T}(\mathcal{X})$ where $\text{ob}(\mathcal{T}(\mathcal{X})) = \tau$ and morphisms are inclusion relations, i.e. for any two open subsets U and V of X , there exists a morphism $V \rightarrow U$ iff $V \subseteq U$. A *presheaf over a topological space* \mathcal{X} is then defined as a functor $P : \mathcal{T}(\mathcal{X})^{op} \rightarrow \mathbf{Sets}$. These inclusion

morphisms $V \rightarrow U$ are mapped under the presheaf to *restriction morphisms* $res_V^U : PU \rightarrow PV$. For each element of $s \in PU$, we will denote the action of the restriction morphism on s as $s|_V$, i.e.:

$$\begin{aligned} res_V^U &: PU \rightarrow PV \\ &:: s \mapsto s|_V \end{aligned} \tag{1.10}$$

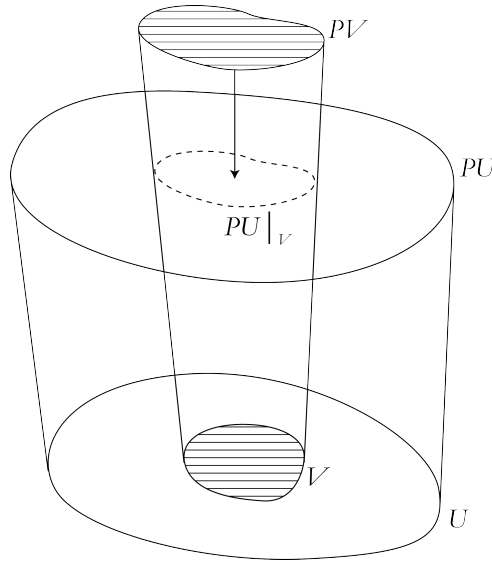


Figure 1.1: Illustration of the restriction morphisms of a presheaf.

Given an open subset U , we will define *sections of P at U* as the elements of the set PU . The intuition is that sections of a presheaf encode *data* over the topological space \mathcal{X} . The existence of the restriction morphisms signifies that data of a smaller subset V can be retrieved from the data of a larger set U (see Fig. 1.1).

Remark 1.13. The notion of *section* (or *cross section*) is usually defined in the literature in terms of a *bundle* (i.e. map) $p : E \rightarrow X$ as maps $\sigma : U \rightarrow E$ such that $U \xrightarrow{\sigma} E \xrightarrow{p} X$ is the inclusion map $U \hookrightarrow X$ [128, 81]. There is, in fact, a one-to-one correspondence between elements of PU and sections of a canonical bundle [128] (see Appendix A for more details).

So far, we have looked at the consistency of data between subsets via restriction maps. We now describe the notion of consistency “across” different open sets. Given a presheaf P over a topological space \mathcal{X} , we say that there is a *gluing* between two

sections $s_U \in PU$ and $s_V \in PV$ of the open sets U and V , or equivalent that s_U and s_V are *locally consistent* or *compatible*, iff:

$$s_U|_{U \cap V} = s_V|_{U \cap V} \quad (1.11)$$

This gluing condition is illustrated in Fig.1.2. The existence of a gluing represents consistency of data at a local level.

In terms of *global consistency*, we want to be able to define a gluing for every pair of open subsets. The notion of *sheaves* encodes this global consistent condition.

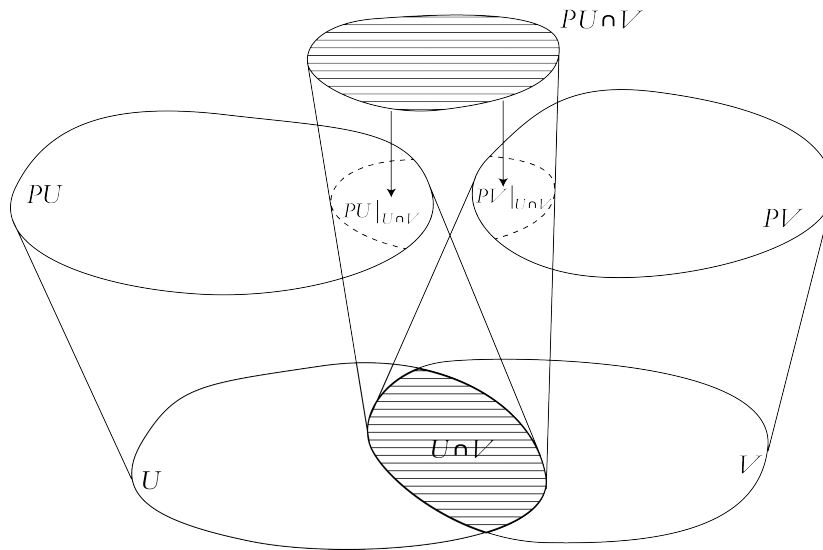


Figure 1.2: Illustration of the general presheaf structure over intersecting sets. If there exists a gluing between two sections in PU and PV , then there will be an intersection between they will coincide in the two dashed regions $PU|_{U \cap V}$ and $PV|_{U \cap V}$.

Definition 1.14. A presheaf $P : \mathcal{T}(\mathcal{X})^{op} \rightarrow \mathbf{Sets}$ is a *sheaf* iff for any covering $\{U_i\}_{i \in I}$ of the space $\mathcal{X} = (X, \tau)$, i.e. $\bigcup_{i \in I} U_i = X$, for every family of pairwise compatible sections $\{s_i \in U_i\}_{i \in I}$, then there exists a section $s \in PX$ such that:

$$s|_{U_i} = s_i \quad (1.12)$$

for all $i \in I$.

Although many had used concepts similar to sheaves before, the exact notion originates from the work of Jean Leray [118], who introduced them to study equa-

tions and transformations from a purely topological perspective – by getting rid of notions he found unnecessary. Subsequent work of Cartan and Serre then exported the ideas from sheaf theory to algebraic geometry [35, 163], which was then made categorical by Grothendieck in the seminal article [85]. The machinery of sheaf theory was then notably used to prove Weil’s conjectures in algebraic geometry [84, 45].

The use of sheaves also arose, somewhat independently, from a logical perspective, notably from the work of Lawvere and Tierney [116, 176]. In particular, the category of sheaves (as well as the category of presheaves) forms a *topos*, which can be associated with a logic [81, 114, 98]. In foundations of mathematics, topos theory has provided alternative (and much simpler) proofs of results from set theory, notably the independence of the Continuum Hypothesis [177] and the Axiom of Choice [70] from the Zermelo-Frænkel axioms. In addition, categories of presheaves and sheaves were found to provide semantics for *intuitionistic logic*, usually referred to as *Kripke-Joyal semantics* [108].

Sheaf theory has recently been applied to topological data analysis [44] and quantum mechanics [6]. We will discuss the latter in more detail in Section 1.2.1.

1.1.3 Monoidal categories

We now turn our attention to another type of category which we will use in Chapter 4, namely *monoidal categories*. As we will see in Section 1.3, these categories are particularly useful in modeling process theories.

Definition 1.15. A *monoidal category* is category \mathcal{C} which is equipped with the structure $(\otimes, I, \alpha, \lambda, \rho)$ where:

- The *tensor product* or *monoidal product* \otimes is a bifunctor: $\otimes : \mathcal{C} \times \mathcal{C} \rightarrow \mathcal{C}$
- $I \in \text{ob}(\mathcal{C})$ is the *unit object*. The morphisms $e : I \rightarrow A$, where $A \in \text{ob}(\mathcal{C})$, will be called the *elements* of A
- The *associator* α is a natural isomorphism with elements $\alpha_{A,B,C} : (A \otimes B) \otimes C \rightarrow A \otimes (B \otimes C)$ for any $A, B, C \in \text{ob}(\mathcal{C})$
- The *left unitor* λ is a natural isomorphism with elements $\lambda_A : I \otimes A \rightarrow A$
- The *right unitor* ρ is a natural isomorphism with elements $\rho_A : A \otimes I \rightarrow A$

such that the following diagrams commute:

$$\begin{array}{ccc}
 (A \otimes I) \otimes B & \xrightarrow{\alpha_{A,I,B}} & A \otimes (I \otimes B) \\
 \rho_A \otimes id_B \searrow & & \swarrow id_A \otimes \lambda_B \\
 & A \otimes B &
 \end{array} \tag{1.13}$$

$$\begin{array}{ccc}
 & (A \otimes B) \otimes (C \otimes D) & \\
 \alpha_{A \otimes B, C, D} \nearrow & & \searrow \alpha_{A, B, C \otimes D} \\
 ((A \otimes B) \otimes C) \otimes D & & A \otimes (B \otimes (C \otimes D)) \\
 \alpha_{A, B, C} \otimes id_D \downarrow & & \uparrow id_A \otimes \alpha_{B, C, D} \\
 (A \otimes (B \otimes C)) \otimes D & \xrightarrow{\alpha_{A, B \otimes C, D}} & A \otimes ((B \otimes C) \otimes D)
 \end{array} \tag{1.14}$$

The equations (1.13) and (1.14) are referred to as the *triangle* and *pentagon* equations respectively.

Example 1.16. Here are some examples of monoidal categories

- a. The category **Sets** is a monoidal category where the tensor product is the standard cartesian product of sets, and the unit is the singleton set $I = \{\star\}$. Since for any sets A, B, C , we have $(A \times B) \times C = A \times (B \times C)$, the associator consists of identity morphisms. In addition, we have $\{\star\} \times A \simeq A$ for any set A and the left unitor is defined as:

$$\lambda_A :: (\star, a) \in I \times A \mapsto a \in A \quad \lambda_A^{-1} :: a \in A \mapsto (\star, a) \in I \times A$$

The right unitor can be defined similarly. Elements of an object A will correspond to the elements a of the set A .

- b. Similarly to the category of **Sets**, the category of sets and relations **Rel** is also monoidal, where the monoidal product is also the cartesian product. The associator is, as in **Sets**, simply consisting of identities and the left and right unitors are defined as:

$$\lambda_A = \{((\star, a), a)\} \subseteq (I \times A) \times A \quad \rho_A = \{((a, \star), a)\} \subseteq (A \times I) \times A \tag{1.15}$$

Elements of an object $A \in ob(\mathbf{Rel})$ will be isomorphic to subsets of A .

- c. The category of finite dimensional vector spaces \mathbf{FdVect} is also a monoidal category. The standard choice of monoidal product is the *tensor product* \otimes . For example, given two vector spaces V and W of dimension n and m respectively, the vector space $V \otimes W$ will have dimension $n \times m$. Furthermore, for each pair of linear maps $\mathbf{A} : U \rightarrow V$ and $\mathbf{B} : W \rightarrow X$ which can be seen as matrices:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{l1} & a_{l2} & \dots & a_{lk} \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{pmatrix}$$

then the matrix corresponding to $\mathbf{A} \otimes \mathbf{B}$ will correspond to:

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1k}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2k}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{l1}\mathbf{B} & a_{l2}\mathbf{B} & \dots & a_{lk}\mathbf{B} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} & a_{12}b_{11} & \dots & a_{11}b_{1m} \\ a_{21}b_{21} & a_{22}b_{21} & \dots & a_{21}b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{l1}b_{n1} & a_{l2}b_{n1} & \dots & a_{lk}b_{nm} \end{pmatrix}$$

It is very convenient to use *string diagrams* to represent objects and morphisms in a monoidal category. Generally speaking, each string diagram corresponds to an equivalence class of morphisms with respect to certain^[2] isomorphisms in \mathcal{C} . The objects of a monoidal category, as well as the identity morphisms, will be represented by wires:

$$\begin{array}{c} | \\ A \end{array}$$

with the exception of the unit object I , which will generally not be represented (i.e., corresponds to the empty wire). Morphisms will be represented by boxes, e.g.:

$$\begin{array}{c} | \\ A \\ \boxed{f} \\ | \\ B \end{array} = f : A \rightarrow B$$

^[2]In general, equality of string diagrams is defined in terms of strict monoidal categories or strict symmetric monoidal categories, see [127] and [149] for respective definitions. We also note that this is reasonable as every (symmetric) monoidal category is monoidally equivalent to a strict (symmetric) monoidal category [127, 198].

As we mentioned before, the identity morphisms will be represented by the wires themselves. Sequential composition of morphisms will be represented by “stacking” boxes, e.g.:

$$\begin{array}{c}
 | A \\
 \boxed{f} \\
 | B \\
 \boxed{g} \\
 | C
 \end{array}
 =
 A \xrightarrow{f} B \xrightarrow{g} C$$

Remark 1.17. Many conventions can be adopted regarding the direction of composition in string diagrams. Here, we will assume that sequential composition happens from top to bottom.

Let us now look at the diagrammatic view of the monoidal structure. We will represent the tensor product of two objects as stacking wires *in parallel*:

$$\begin{array}{c}
 | \\
 | A \\
 | \\
 | B \\
 |
 \end{array}$$

Similarly, the action of the tensor product on morphisms will also be represented by concatenating the boxes in parallel, e.g.:

$$\begin{array}{c}
 | A \quad | C \\
 \boxed{f} \quad \boxed{g} \\
 | B \quad | D
 \end{array}$$

In particular, from the associator being a natural isomorphism coupled with the pentagon equation (1.14), it is not necessary to specify in which order the objects are being tensored as they will all lead to isomorphic objects (and therefore the same string diagram); consequently, we also do not need to represent morphisms $\alpha_{A,B,C}$. Similarly, the representation of the unit object I as the empty wire is motivated by the existence of the left and right unitors, as well as the triangle equation (1.13), i.e.:

$$\begin{array}{c}
 | \\
 | A \\
 |
 \end{array}$$

would represent objects A , $I \otimes A$ and $A \otimes I$ alike.

In addition, *elements* $e : I \rightarrow A$ will be denoted as triangles in string diagrams,

namely:

$$\begin{array}{c} \triangle \\ \uparrow \\ A \end{array} \begin{array}{c} e \\ \\ \end{array} = e : I \rightarrow A$$

Monoidal categories can be seen as process theories. Indeed, taking objects of the category to be systems or entities and morphisms to be processes, we can see the sequential composition of morphisms as the evolution of the entities through the sequence of processes. Similarly, the tensor product of morphisms will correspond to the execution of independent processes in parallel. Then, the string diagrams are graphical representations of the interactions of different systems under the action of some processes.

Symmetric monoidal categories So far, the definition of a monoidal category only describes a very general process theory. In order to describe process theories satisfying some extra properties, we define more and more specialised monoidal categories by adding additional structure or axioms.

We start by defining a *symmetric monoidal category*, which is a monoidal category \mathcal{C} equipped with a natural isomorphism with elements $\sigma_{A,B} : A \otimes B \rightarrow B \otimes A$ such that :

$$\sigma_{B,A} \circ \sigma_{A,B} = id_{A \otimes B} \tag{1.16}$$

In terms of string diagrams, we represent the isomorphisms $\sigma_{A,B}$ as follows:



In addition, by naturality of σ , this implies that “morphisms slides through the crossings”, i.e.:

$$\begin{array}{ccc} \begin{array}{c} A \\ | \\ \boxed{f} \\ | \\ C \end{array} \begin{array}{c} B \\ | \\ \boxed{g} \\ | \\ D \end{array} & = & \begin{array}{c} A \\ \diagdown \quad \diagup \\ B \quad A \end{array} \\ \begin{array}{c} C \\ | \\ \boxed{g} \\ | \\ D \end{array} \begin{array}{c} D \\ | \\ \boxed{f} \\ | \\ C \end{array} & & \begin{array}{c} B \\ | \\ \boxed{g} \\ | \\ D \end{array} \begin{array}{c} A \\ | \\ \boxed{f} \\ | \\ C \end{array} \end{array} \iff \begin{array}{ccc} A \otimes B & \xrightarrow{f \otimes g} & C \otimes D \\ \sigma_{A,B} \downarrow & & \downarrow \sigma_{C,D} \\ B \otimes A & \xrightarrow{g \otimes f} & C \otimes D \end{array} \tag{1.17}$$

As these morphisms are isomorphisms, two diagrams are equivalent in a symmetric

monoidal category if they can be transformed into each other by crossing or uncrossing wires.

Dual objects We now introduce another property that monoidal categories can have which will become very useful in Section 1.3, namely dual objects.

Definition 1.18. (Dual objects) An object R in a monoidal category \mathcal{C} has a *left-dual* $L \in \text{ob}(\mathcal{C})$, or equivalently L has the *right-dual* R iff there exists morphisms $\eta : I \rightarrow R \otimes L$, known as the *unit*, and $\epsilon : L \otimes R \rightarrow I$, known as the *counit*, such that the following diagrams commute:

$$\begin{array}{ccc}
 L & \xrightarrow{\rho_R^{-1}} & L \otimes I \xrightarrow{id_L \otimes \eta} L \otimes (R \otimes L) \\
 id_L \downarrow & & \downarrow \alpha_{L,R,L}^{-1} \\
 L & \xleftarrow{\lambda_L} & I \otimes L \xleftarrow{\epsilon \otimes id_L} (L \otimes R) \otimes L
 \end{array} \tag{1.18}$$

$$\begin{array}{ccc}
 R & \xrightarrow{\lambda_L^{-1}} & I \otimes R \xrightarrow{\eta \otimes id_R} (R \otimes L) \otimes R \\
 id_R \downarrow & & \downarrow \alpha_{R,L,R} \\
 R & \xleftarrow{\rho_R} & R \otimes I \xleftarrow{id_R \otimes \epsilon} R \otimes (L \otimes R)
 \end{array} \tag{1.19}$$

In terms of string diagrams, it is useful to represent dual objects by decorating the wires with arrows going in the opposite direction, e.g.:

$$L = \begin{array}{c} | \\ \downarrow \\ L \end{array} \quad R = \begin{array}{c} | \\ \uparrow \\ R \end{array} \tag{1.20}$$

In addition, by representing the morphisms η and ϵ as:

$$\eta = \begin{array}{c} \text{---} \\ \uparrow R \quad \downarrow L \\ \text{---} \end{array} \quad \epsilon = \begin{array}{c} L \downarrow \quad \uparrow R \\ \text{---} \\ \text{---} \end{array} \tag{1.21}$$

If a monoidal category \mathcal{C} is both symmetric and rigid, it is said to be a *compact-closed category*.

Dagger structures The dualising functor is not the only interesting example of a functor $\mathcal{C} \rightarrow \mathcal{C}^{op}$. Indeed, the notion of *dagger functor* $\dagger : \mathcal{C} \rightarrow \mathcal{C}^{op}$ will become useful in the description of Hilbert spaces to encode the notion of inner products.

Definition 1.21. For any category \mathcal{C} , we define a *dagger functor* as a functor $\dagger : \mathcal{C} \rightarrow \mathcal{C}^{op}$ such that for any morphism f in \mathcal{C} , $(f^\dagger)^\dagger = f$. A category with a dagger functor is a *dagger category*. Respectively, a monoidal category endowed with a dagger functor is called a *dagger monoidal category*.

Example 1.22. We finally introduce the category of Hilbert spaces \mathbf{Hilb} and its subcategory \mathbf{FdHilb} of finite-dimensional Hilbert spaces. In these categories, the objects are Hilbert spaces \mathcal{H} , i.e. complex vector spaces equipped with an inner product $\langle _ | _ \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{C}$ such that $d(x, y) = \sqrt{\langle x - y | x - y \rangle_{\mathcal{H}}}$ is a metric. The morphisms in \mathbf{Hilb} and \mathbf{FdHilb} will be taken to be *bounded* (or equivalently continuous with respect to the topology induced by the metric) linear maps between Hilbert spaces. Then, we define the action of \dagger on objects to return the same Hilbert space, i.e. $\mathcal{H}^\dagger = \mathcal{H}$ for any $\mathcal{H} \in \text{ob}(\mathbf{Hilb})$ (resp. \mathbf{FdHilb}) and for any bounded linear map $f : \mathcal{H} \rightarrow \mathcal{K}$, we take $f^\dagger : \mathcal{K} \rightarrow \mathcal{H}$ to be the unique map such that:

$$\langle f(\mathbf{v}) | \mathbf{w} \rangle_{\mathcal{K}} = \langle \mathbf{v} | f^\dagger(\mathbf{w}) \rangle_{\mathcal{H}} \quad (1.26)$$

for any $\mathbf{v} \in \mathcal{H}$ and $\mathbf{w} \in \mathcal{K}$. Treating morphisms in \mathbf{FdHilb} as complex matrices, taking the dagger correspond to taking the Hermitian conjugate of the matrix, i.e. given $\mathbf{A} : \mathcal{H} \rightarrow \mathcal{K}$, $\mathbf{A}^\dagger = (\mathbf{A}^T)^* = (\mathbf{A}^*)^T$.

A similar definition of a dagger functor can also be obtained for \mathbf{FdVect} , but will then depend on a choice of inner product.

Finally, in monoidal categories coming with both a dualising functor and a dagger functor, we may require some additional conditions on the interaction between the two structures.

Definition 1.23. A compact closed category (i.e. symmetric and autonomous) with a dagger functor is a *dagger-compact category* whenever the following hold for any

pair of duals (L, R) :

$$\left(\begin{array}{c} L \downarrow \\ \text{---} \\ \uparrow R \end{array} \right)^\dagger = \begin{array}{c} R \uparrow \\ \text{---} \\ L \downarrow \\ \text{---} \\ \uparrow R \end{array} \quad \left(\begin{array}{c} R \uparrow \\ \text{---} \\ \downarrow L \end{array} \right)^\dagger = \begin{array}{c} R \downarrow \\ \text{---} \\ L \uparrow \\ \text{---} \\ \downarrow L \end{array} \quad (1.27)$$

Example 1.24. The categories **Hilb** and **FdHilb** can be seen to be dagger-compact by taking the duals to be the same as the ones defined earlier for vector spaces and the dagger functor defined above.

1.2 Describing Quantum Correlations

The aim of this section is to introduce the concept of quantum contextuality, and in particular the framework of the sheaf-theoretic contextuality (introduced in details in Section 1.2.2), which will be a recurrent theme of the work described in Part II and III. The framework of Contextuality-by-Default is also introduced in Section 1.2.3, which will be widely used in Chapter 3.

The inherent probabilistic nature of quantum mechanics has been a longstanding source of debate. Namely, is nature non-deterministic, or is the apparent randomness due to our lack of knowledge about the observed system? This question has led to the development of theory-agnostic descriptions of observations from quantum systems. That is, only assuming classical probability theory, is it possible to describe the observed statistics? Or are the statistical correlations intrinsically non-classical (i.e. different from probabilistic classical physical systems)? These questions are answered by studying the *contextuality* of quantum systems.

We first introduce the standard formalisms of contextuality (Section 1.2.1), and then describe its categorical equivalent and its various extensions (Section 1.2.2). In Section 1.2.3, we introduce an alternative framework to the one of Section 1.2.2.

In terms of notation, we will also use the standard *Dirac notation* where vectors in a Hilbert space will be denoted as $|\psi\rangle \in \mathcal{H}$, their Hermitian conjugate will be denoted as $\langle\psi|$, and the inner product of two vectors $|\psi\rangle, |\phi\rangle \in \mathcal{H}$ will be denoted as $\langle\phi|\psi\rangle$.

1.2.1 Contextuality

The initial criticism of Einstein, Podolsky, and Rosen [58] was that the probabilistic nature of quantum mechanics can only be due to the *incompleteness* of quantum theory. This is known as the EPR paradox. Hence, by “completing” the description of the system with additional (unobserved) variables, one could obtain a deterministic system from which we can recover the observed statistics [58]. The main argument from [58] is that any physical theory should satisfy *realism*, i.e. every physical quantity, such as the position or the momentum of a particle, should possess a definite value at any given time which should not depend on whether it is observed or not.

Non-locality

The first and most widely known counterargument of the EPR paradox is attributed to John Bell [22]. In reality, Bell’s theorem uses an assumption not made explicit in [58], namely that spatially separated systems cannot influence each other. This requirement is known as *no-signalling*.

We here describe an operational view of Bell’s theorem due to Fine [63]. Let’s consider an experiment such that a party Charlie prepares a bipartite state $|\Psi\rangle$ where one subsystem is sent to Alice, the other to Bob, such that Alice and Bob are assumed to be so far away that they cannot influence each other in the time frame of the experiment (see Fig. 1.3). Then, both Alice and Bob randomly (and independently) choose to measure a physical quantity on their subsystems, say in the respective sets $\{a, a'\}$ and $\{b, b'\}$. Finally, each physical quantity will take values in ± 1 .

From the realism condition, we will require that, given a hidden-variable $\lambda \in \Lambda$, the values of physical quantities a, a', b, b' are uniquely (and deterministically) determined. In addition, from the no-signalling condition, we will require that the outcomes of a, a' will not depend on the choice of Bob, and respectively, the outcomes of b or b' are independent of Alice’s choice. Therefore, using both of these requirements, we can define functions $A : \{a, a'\} \times \Lambda \rightarrow \{\pm 1\}$ and $B : \{b, b'\} \times \Lambda \rightarrow \{\pm 1\}$ which associate the value of the physical quantities accessible from Alice and Bob’s

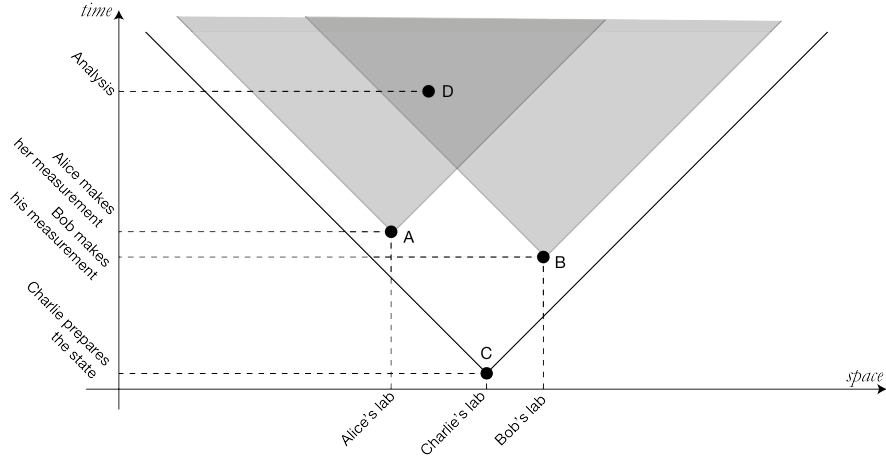


Figure 1.3: Causal diagram of a Bell experiment. Events are represented as dot and future light-cones are represented as triangles.

measurements given the value of a hidden variable Λ ^[3].

Using these conditions, it is possible to show that:

$$|\langle ab \rangle + \langle ab' \rangle + \langle a'b \rangle - \langle a'b' \rangle| \leq 2 \quad (1.28)$$

The proof of this can be found in Appendix B. This inequality is widely known as the *CHSH inequality* (for Clauser-Horne-Shimony-Holt who first proved it) [40]. Generally speaking, any inequality that provides a sufficient condition for the existence of a hidden variable model is known as a *Bell inequality*^[4].

However, we know that quantum theory predicts violations of such Bell inequalities, and this has now been verified experimentally [91, 78, 164, 159, 172]. One example of such violation of (1.28) can be achieved by taking the state:

$$|\Psi\rangle = \frac{1}{\sqrt{2}} (|0\rangle \otimes |1\rangle - |0\rangle \otimes |1\rangle)$$

as the state prepared by Charlie, and a, a', b, b' being one-qubit measurements along the respective basis [141]:

^[3]Note that we can assume without loss of generality that we only have a single hidden variable Λ . The case of multiple hidden variables $\Lambda_1, \Lambda_2, \dots$ can be reduced to a single hidden-variable model by taking the joint distributions over $\prod_i \Lambda_i$.

^[4]The original inequality proved by Bell in [22] corresponds to a different experiment in which it is harder to obtain a violation [40].

Measurement	Outcome -1	Outcome $+1$
a	$ 0\rangle$	$ 1\rangle$
a'	$\frac{1}{\sqrt{2}}(0\rangle + 1\rangle)$	$\frac{1}{\sqrt{2}}(0\rangle - 1\rangle)$
b	$\cos\left(\frac{-\pi}{8}\right) 0\rangle + \sin\left(\frac{-\pi}{8}\right) 1\rangle$	$\cos\left(\frac{3\pi}{8}\right) 0\rangle + \sin\left(\frac{3\pi}{8}\right) 1\rangle$
b'	$\cos\left(\frac{\pi}{8}\right) 0\rangle + \sin\left(\frac{\pi}{8}\right) 1\rangle$	$\cos\left(\frac{5\pi}{8}\right) 0\rangle + \sin\left(\frac{5\pi}{8}\right) 1\rangle$

From these sets of measurements, it can be shown that the observed probability distributions are given by:

	$(-1, -1)$	$(-1, +1)$	$(+1, -1)$	$(+1, +1)$
(a, b)	$\frac{1}{2} \sin^2\left(\frac{\pi}{8}\right)$	$\frac{1}{2} \cos^2\left(\frac{\pi}{8}\right)$	$\frac{1}{2} \cos^2\left(\frac{\pi}{8}\right)$	$\frac{1}{2} \sin^2\left(\frac{\pi}{8}\right)$
(a, b')	$\frac{1}{2} \sin^2\left(\frac{\pi}{8}\right)$	$\frac{1}{2} \cos^2\left(\frac{\pi}{8}\right)$	$\frac{1}{2} \cos^2\left(\frac{\pi}{8}\right)$	$\frac{1}{2} \sin^2\left(\frac{\pi}{8}\right)$
(a', b)	$\frac{1}{4} \left(1 + \frac{1}{\sqrt{2}}\right)$	$\frac{1}{4} \left(1 - \frac{1}{\sqrt{2}}\right)$	$\frac{1}{4}$	$\frac{1}{4}$
(a', b')	$\frac{1}{4} \left(1 - \frac{1}{\sqrt{2}}\right)$	$\frac{1}{4} \left(1 + \frac{1}{\sqrt{2}}\right)$	$\frac{1}{4}$	$\frac{1}{4}$

Hence, by some simple calculations, it is possible to obtain the expectation values of the product variables ab , ab' , $a'b$, and $a'b'$ as:

$\langle ab \rangle$	$\langle ab' \rangle$	$\langle a'b \rangle$	$\langle a'b' \rangle$
$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	$\frac{-1}{\sqrt{2}}$

Which leads to:

$$|\langle ab \rangle + \langle ab' \rangle + \langle a'b \rangle - \langle a'b' \rangle| = 2\sqrt{2} > 2 \quad (1.29)$$

The implications of the violation of (1.28) is that the statistics of quantum systems cannot be explained by a model consistent with realism (i.e. all physical quantities have values in a system at a given time) and locality (i.e. an event can only influence other events in its future light-cone). Which is the correct assumption to drop, i.e. realism or locality, is still highly debated in quantum foundations.

A theory without locality would imply that the physical property of a system (e.g. its position) can instantly and non-locally be altered. The leading example of such interpretation is known as *Bohmian mechanics* [27, 28]. We should also note that, even in non-local theories, *information* cannot travel faster than light. For example, in the previously described experiment, Alice cannot infer which measurement Bob has done from her observed local statistics, and conversely.

On the other hand, non-realistic but local theories can also be formulated. Examples of such interpretations are the so-called *Copenhagen interpretation* [29], which

promotes the idea of the “collapse of the wavefunction” upon measurement, or *QBism* [37] in which a quantum state is not the “reality” of the system, but correspond to a more subjective description of it.

Contextuality

The proof of the non-existence of hidden variables in quantum mechanics above highly depends on the locality or no-signalling condition, which restricts the situations it can describe. In [107], the so-called *Kochen-Specker theorem* proposed a more general criterion for the non-existence of hidden variables. Indeed, instead of relying on spatially separated measurements, we only require that measurements be done “at the same time”, meaning that we can know the values of the measurements simultaneously. For example, if the value of an observable A is found to be 3, then the value of the operator A^2 is automatically known to be 9. In the standard quantum mechanics formalism, this condition is expressed as having commuting observables, i.e. A and B as compatible iff $AB = BA$. If two observables are compatible we should be able to observe their joint statistics, whereas it does not make sense to talk about the joint statistics of observables which are not compatible. A system will then be said to be non-contextual if we can extend the system to one where all of the observables are compatible, i.e. a system in which the values of all the observables can be known at the same time. This extended system can be seen as a hidden-variable model of the system. The Kochen-Specker theorem [107] then states that this is not possible for the observables of quantum mechanics. We now describe their proof, as well as subsequent proofs, of quantum contextuality.

The Kochen-Specker argument was originally abstractly formulated in terms of partial algebra as follows. We start with a set of observables \mathcal{O} endowed with a *co-measurability relation* $\mathfrak{R} \subseteq \mathcal{O} \times \mathcal{O}$; namely, for two operators $A, B \in \mathcal{O}$, we have $A \mathfrak{R} B$ iff A and B are compatible. This co-measurability relation is required to be reflexive (i.e. every observable is co-measurable with itself) and symmetric (i.e. if A is co-measurable with B , then B is co-measurable with A). In addition, it is also desirable that, if A and B are both compatible with an observable C , then any function^[5] of A and B will remain compatible with C . Formally, the co-measurability relation corresponds to a *partial algebra* defined as follows.

^[5]Strictly speaking, we mean a Borel function here.

Definition 1.25 (Partial algebra). A *partial algebra* over a field K is a tuple (X, \mathfrak{R}) where X is a set endowed with addition, multiplication, and scalar multiplication over K , and \mathfrak{R} is a binary relation $\mathfrak{R} \subseteq X \times X$ such that:

- a. \mathfrak{R} is symmetric and reflexive;
- b. There exists an element $1 \in X$ such that $A \mathfrak{R} 1$ for all $A \in X$
- c. For any $A_1, A_2, A_3 \in X$ such that $A_i \mathfrak{R} A_j$ for all $i, j \in \{1, 2, 3\}$ and $\alpha \in K$:

$$(A_1 + A_2) \mathfrak{R} A_3 \quad (1.30)$$

$$A_1 \cdot A_2 \mathfrak{R} A_3 \quad (1.31)$$

$$\alpha A_1 \mathfrak{R} A_2 \quad (1.32)$$

- d. For any A_1, A_2, A_3 such that $A_i \mathfrak{R} A_j$ for all $i, j \in \{1, 2, 3\}$, the polynomials over A_1, A_2, A_3 form a commutative algebra.

In particular, in the case where the field K is the field \mathbb{F}_2 , we talk of partial Boolean algebras.

In addition, we can define morphisms between partial algebras as follows.

Definition 1.26. A *partial algebra homomorphism* is a map $h : (X, \mathfrak{R}_X) \rightarrow (Y, \mathfrak{R}_Y)$ between two partial algebras over the same field K such that:

$$a \mathfrak{R}_X b \implies h(a) \mathfrak{R}_Y h(b) \quad (1.33)$$

$$h(\alpha a + \beta b) = \alpha h(a) + \beta h(b) \quad \forall \alpha, \beta \in K \quad (1.34)$$

$$h(a \cdot b) = h(a) \cdot h(b) \quad (1.35)$$

$$h(1) = 1 \quad (1.36)$$

Therefore, if there exists a homomorphism $h : (X, \mathfrak{R}_X) \rightarrow (Y, \mathfrak{R}_Y)$, then the theory over observables in X can be simulated using observables in Y such that the functional relations between the observables of X are preserved.

Then, a partial algebra \mathfrak{R} is said to be *non-contextual* iff there exists a partial algebra homomorphism $h : \mathfrak{R} \rightarrow \mathfrak{A}$ where \mathfrak{A} is a *total* (commutative) algebra. Accordingly, a partial Boolean algebra is non-contextual iff there is a partial algebra homomorphism into a Boolean algebra.

This is indeed the case for classical mechanics where observables are functions $f : \Omega \rightarrow \mathbb{R}$ with Ω being the set of possible states of the system. This indeed forms a commutative algebra over \mathbb{R} , where all of the observables are co-measurable, and we have the following:

$$\begin{aligned} 1 : \Omega \rightarrow \mathbb{R} &::= \omega \mapsto 1 \\ f + g : \Omega \rightarrow \mathbb{R} &::= \omega \mapsto f(\omega) + g(\omega) \\ f \cdot g : \Omega \rightarrow \mathbb{R} &::= \omega \mapsto f(\omega) \times g(\omega) \\ \alpha f : \Omega \rightarrow \mathbb{R} &::= \omega \mapsto \alpha f(\omega) \end{aligned}$$

On the other hand, the Kochen-Specker theorem states that quantum mechanical observables are *contextual*, i.e. that there is no homomorphism to a total commutative algebra. Indeed, quantum observables are represented as self-adjoint operators on a Hilbert space \mathcal{H} . Addition, multiplication, and scalar multiplication of operators is defined as the standard matrix operations. As mentioned previously, the co-measurability \mathfrak{R} is defined such that $A \mathfrak{R} B$ iff $AB = BA$. This structure forms a partial algebra. The obtained partial algebra is not total, as in general, not all observables will commute. Therefore, a set of quantum observables \mathcal{O} admits a hidden variable model iff there exists a homomorphism $h : (\mathcal{O}, \mathfrak{R}) \rightarrow (\mathcal{O}', \mathfrak{R}')$ where $(\mathcal{O}', \mathfrak{R}')$ is the partial algebra associated with a total (commutative) algebra.

In the original article [107], the proof of contextuality for quantum theory was achieved by looking at a set of 117 observables in a 3-dimensional Hilbert space (this could represent the angular momentum of a single particle along 117 different directions). Later on, simpler proofs of quantum contextuality have been proposed, using smaller sets of observables, and provided less involved geometric arguments [143, 132, 34].

In addition, a major flaw in the original proof of [107] is that it is not easily checked experimentally. In [106], Klyachko, Can, Binicioğlu and Shumovsky (KCBS) provided a contextuality proof on a 3-dimensional quantum system by deriving a non-contextual inequality, in the same vein as the CHSH inequality for non-locality, and showing its violation by fixing a set of 5 projection operators and the state which is being measured.

To see this, we start with the derivation of the classical bound. Suppose that we have 5 observables $\{A_k\}_{k=1,\dots,5}$ such that A_k and $A_{k \oplus_5 1}$ are co-measurable for all k

(where \oplus_5 denotes the addition modulo 5). Now, suppose that all of the A_k 's take value in ± 1 , therefore, the products $A_k A_{k\oplus 5 1}$ test whether their values are correlated (i.e. $A_k A_{k\oplus 5 1} = 1$) or anticorrelated (i.e. $A_k A_{k\oplus 5 1} = -1$). Then, if there exists a hidden-variable model, all of the A_k gets assigned a value, regardless of which pair of observables is measured (see Table 1.1). Moreover, since there is an odd number of pairs $(A_k, A_{k\oplus 5 1})$, then the number of anticorrelated pairs has to be even in a global assignment of values, and is at most 4. Hence, this gives the KCBS inequality:

$$\sum_{k=1}^5 \langle A_k A_{k\oplus 5 1} \rangle \geq -3 \quad (1.37)$$

It turns out that the assignment of Table 1.1 (seen as a deterministic hidden-variable model) saturates this inequality.

	A_1	A_2	A_3	A_4	A_5
Value	+1	-1	+1	+1	-1

Table 1.1: Example of a total assignment of values to the observables A_k in the KCBS experiment.

Now, we will describe a specific instance of such an experiment demonstrating contextuality in 3-dimensional quantum systems, which is taken from [33, 10]. Starting from the 5 states (up to normalisation factors):

$$|v_1\rangle \propto |0\rangle + \sqrt{\cos\left(\frac{\pi}{5}\right)} |2\rangle \quad (1.38)$$

$$|v_2\rangle \propto \cos\left(\frac{4\pi}{5}\right) |0\rangle + \sin\left(\frac{4\pi}{5}\right) |1\rangle + \sqrt{\cos\left(\frac{\pi}{5}\right)} |2\rangle \quad (1.39)$$

$$|v_3\rangle \propto \cos\left(\frac{2\pi}{5}\right) |0\rangle - \sin\left(\frac{2\pi}{5}\right) |1\rangle + \sqrt{\cos\left(\frac{\pi}{5}\right)} |2\rangle \quad (1.40)$$

$$|v_4\rangle \propto \cos\left(\frac{2\pi}{5}\right) |0\rangle + \sin\left(\frac{2\pi}{5}\right) |1\rangle + \sqrt{\cos\left(\frac{\pi}{5}\right)} |2\rangle \quad (1.41)$$

$$|v_5\rangle \propto \cos\left(\frac{4\pi}{5}\right) |0\rangle - \sin\left(\frac{4\pi}{5}\right) |1\rangle + \sqrt{\cos\left(\frac{\pi}{5}\right)} |2\rangle \quad (1.42)$$

It can be checked that these states are pairwise orthogonal, i.e. they satisfy $\langle v_k | v_{k\oplus 5 1} \rangle = 0$ for each k . These states give a set of projections operators with eigenvalues (i.e. outcomes) ± 1 , namely:

$$P_k = 2 |v_k\rangle \langle v_k| - \mathbb{I} \quad (1.43)$$

where \mathbb{I} is the identity. We can then check that the pairs of projectors P_k and $P_{k\oplus_5 1}$ commute since:

$$P_k P_{k\oplus_5 1} = P_{k\oplus_5 1} P_k = -2 |v_k\rangle \langle v_k| - 2 |v_{k\oplus_5 1}\rangle \langle v_{k\oplus_5 1}| + \mathbb{I}$$

using the fact that the pairs $|v_k\rangle$ and $|v_{k\oplus_5 1}\rangle$ are orthogonal. We recall that this means that changing the order of the projections will not change the values of the individual observables. Now, taking the state to be measured to be the state $|\psi\rangle = |2\rangle$, it can be shown that the expectation value $\langle P_k P_{k\oplus_5 1} \rangle$ is:

$$\langle P_k P_{k\oplus_5 1} \rangle = \frac{1 - 3 \cos\left(\frac{\pi}{5}\right)}{2 \cos^2\left(\frac{\pi}{10}\right)} \quad (1.44)$$

for each $k = 1, \dots, 5$. Therefore leading to the violation of the KCBS inequality (1.37):

$$\sum_{k=1}^5 \langle A_k A_{k\oplus_5 1} \rangle = 5 \frac{1 - 3 \cos\left(\frac{\pi}{5}\right)}{2 \cos^2\left(\frac{\pi}{10}\right)} \simeq -3.944 < -3 \quad (1.45)$$

The advantages of this proof is that it provides clear experiments which needs to be performed for showing contextuality of quantum mechanics, and the inequality derived is minimal in terms of number of observables and dimension of the quantum system [106]. In addition, we should emphasize that this proof of contextuality *does not depend on locality assumptions*, as measurements are done on a single system. This then shows that contextuality is strictly more general than non-locality.

The KCBS inequality was generalised for n -dimensional quantum systems with $n \geq 3$, by considering n observables $\{P_i\}_{i=1, \dots, n}$, where the only compatible measurements are $P_i, P_{i\oplus_n 1}$, where \oplus_n is the addition modulo n [13]. Then, the KCBS inequality arises as the special case $n = 3$, whilst the CHSH inequality corresponds to the case $n = 4$.

Contextuality and quantum computations

Contextuality provides a fundamental distinction between classical and quantum theories and has also been shown to be an essential resource in quantum computing. It has famously been demonstrated that quantum systems can solve computational problems exponentially faster than any known classical algorithm, such as

factoring [166] or simulation of physical systems [16]. Where the advantage comes from has historically been unclear. Recent studies have shown that contextuality is a crucial ingredient for obtaining a quantum advantage, more so than superposition or entanglement, which can be efficiently simulated using classical computers [169, 140].

One of the first demonstrations of the role of contextuality in computation relates to fault-tolerant stabiliser quantum computing.

Fault tolerance is vital to achieve reliable computation on real quantum computers. One of the promising avenues to achieve quantum fault tolerance relies on *stabiliser codes*, where a specific set of measurements (usually generalisations of the Pauli gates) is used to correct noise introduced in a quantum circuit. However, these gates or measurements are part of the Clifford group, which can only generate circuits that are simulable on classical computers [82]. The full power of quantum computations can be achieved from *magic state distillation*.

For magic state distillation, we start from an input state ρ , which can be noisy, and aim at distilling it into a target “magic state” $|m\rangle$ using stabiliser measurements on some subsystem. This target state is defined so that non-Clifford gates can be performed using it.

Now, not all initial states ρ can be distilled into a magic state $|m\rangle$. In [94], the authors showed that the set of states that can be distilled into magic states are precisely the ones that can exhibit contextuality. Since quantum circuits using only Clifford gates are efficiently simulable using classical resources, this result shows that contextuality is essential to obtain a quantum advantage.

Contextuality has also been studied from a resource theoretical point of view. One result is that the amount of contextuality of a system cannot increase with classical operations such that classical pre- and post-processing, classical control over measurements or probabilistic mixing of experiments [3, 48, 12, 190]. This result implies that any computational advantage coming from contextuality, e.g. in magic state distillation, cannot be created from classical operations. Using quantum systems is, therefore, necessary to obtain a quantum advantage.

1.2.2 The sheaf-theoretic view of contextuality

In [6], the authors showed that contextuality corresponds to the impossibility of finding a global section given a consistent family of local sections of a presheaf.

In this framework, the possible local measurements form a set X , which will become the base space of our presheaves (with suitable topology). We then impose a compatibility relation on X , which, in turn, gives us a cover of this space. This compatibility relation corresponds to the co-measurability relation described in Section 1.2.1.

Example 1.27. Let's consider the standard (2,2,2)-Bell scenario consisting of 2 parties, each choosing between 2 measurements, and each measurement having two possible outcomes. The set of possible measurements is $X = \{a_1, a_2, b_1, b_2\}$ and we will denote as $I_A = \{a_i\}_{i=1,2}$ the set of measurements available to Alice, and $I_B = \{b_i\}_{i=1,2}$ the set of measurements available to Bob. Alice's measurements in I_A are compatible with all of Bob's in I_B . However, the measurements a_1 and a_2 are incompatible, as they cannot be performed simultaneously, and similarly for Bob's measurements.

Each of these measurements comes with a set of possible outcomes O ^[6]. Then, given a set of compatible measurements U , an event associates outcomes with the measurements selected in U . An event is, therefore, modelled as a function:

$$s : U \rightarrow O$$

Example 1.28. In the (2,2,2)-Bell scenario, if Alice chooses to perform the measurement a_1 and obtains outcome $x \in O$, and Bob the measurement b_2 and obtains the outcome $y \in O$, then the event could be represented as the function:

$$s : U \rightarrow O :: a_1 \mapsto x; b_2 \mapsto y$$

Presheaves and empirical models

Formally speaking, these functions are modelled as the *presheaf of events*. The presheaf of events is defined as $\mathcal{E} : \mathcal{P}(X)^{op} \rightarrow \mathbf{Sets}$, where the morphisms in $\mathcal{P}(X)$ are inclusion relations. In other words, we are taking a presheaf over the set of measurements X endowed with the *discrete topology*.

^[6]Without loss of generality, we can assume that O is the same for any choice of measurement.

The action of this *presheaf* on objects U gives the set of all possible *assignments* or *functions* $s : U \rightarrow O$. The action on morphisms $U \xrightarrow{\subseteq} V$ in \mathcal{C} , gives us the *restrictions* of these assignments, namely:

$$\begin{array}{ccc} \mathcal{E}(U \subseteq V) : & \mathcal{E}(V) & \rightarrow & \mathcal{E}(U) \\ s : & V \rightarrow O & \mapsto & s|_U : U \rightarrow O \\ & v \in V \mapsto o_v \in O & & v \in U \mapsto o_v \in O \end{array} \quad (1.46)$$

In quantum mechanics, however, the outcomes of measurements are not generally deterministic, so instead of looking at events, it is more relevant to look at *the probability distributions* over all of the possible events. Therefore, we post-compose the event presheaf \mathcal{E} with the distribution monad $\mathcal{D}_{\mathbb{R}_+} : \mathbf{Sets} \rightarrow \mathbf{Sets}$ as defined in Section 1.1. The obtained functor is once again a presheaf.

In a given experiment, we will not observe *all* of the possible probability distributions for each set of co-measurable measurements, but instead, we will see only a *single* probability distribution per global measurement choice, which will correspond to the *observed probability distribution*. Hence, in terms of the presheaf $\mathcal{D}_{\mathbb{R}_+} \mathcal{E}$, this means that, when selecting a set of measurements $U \subseteq X$ to perform, we will only observe a single section $e_U \in \mathcal{D}_{\mathbb{R}_+} \mathcal{E}(U)$.

Similarly, we can only access the probability distributions of specific combinations of compatible measurements in a given quantum experiment. For example, suppose Alice can either measure a_1 or a_2 . In that case, we cannot observe the joint statistics of a_1 and a_2 as these measurements cannot be performed simultaneously. So, instead of looking at sections of $\mathcal{D}_{\mathbb{R}_+} \mathcal{E}(U)$ for each of the subsets $U \subseteq X$, we will instead consider a collection $\mathcal{U} = \{U_i\}_{i \in I}$, such that for each of the collections of $U \in \mathcal{U}$, the elements of U correspond to compatible measurements. Without loss of generality, we will moreover assume that the set \mathcal{U} is a cover of the space X , i.e. $\bigcup_{U \in \mathcal{U}} U = X$, so that all of the measurements are possible. This gives rise to the notion of *measurement scenario*.

Definition 1.29 (Measurement scenario). A *measurement scenario* will consist on a tuple $(\mathcal{X}, \mathcal{U})$, where \mathcal{X} is a topological space and \mathcal{U} is an (open) cover of X .

We then define the data of an experiment as follows.

	(0, 0)	(0, 1)	(1, 0)	(1, 1)
(a_1, b_1)	1/2	0	0	1/2
(a_1, b_2)	3/8	1/8	1/8	3/8
(a_2, b_1)	3/8	1/8	1/8	3/8
(a_2, b_2)	1/8	3/8	3/8	1/8

Table 1.2: An empirical model for the (2,2,2)-Bell scenario

Definition 1.30 (Empirical model). Given measurement scenario $(\mathcal{X}, \mathcal{U})$, we define an *empirical model* as a set of sections $e = \{e_U \in \mathcal{D}_{\mathbb{R}_+} \mathcal{E}(U) \mid U \in \mathcal{U}\}$ of the presheaf $\mathcal{D}_{\mathbb{R}_+} \mathcal{E}(U) : \mathcal{T}(\mathcal{X})^{op} \rightarrow \mathbf{Sets}$, where \mathcal{E} can be any presheaf of events.

Example 1.31. In a (2,2,2)-Bell scenario as described previously, we would have $X = \{a_1, a_2, b_1, b_2\}$, with associated cover $\mathcal{U} = \{\{a_1, b_1\}, \{a_1, b_2\}, \{a_2, b_1\}, \{a_2, b_2\}\}$. Then, an empirical model could be represented as in Table 1.2 where each of the rows is labelled by the choice of measurements $U \in \mathcal{U}$ and correspond to the selected section of $\mathcal{D}_{\mathbb{R}_+} \mathcal{E}(U)$. More specifically, the cell at the intersection of the row labelled by $\{a_i, b_j\}$ and column labelled by $(o_k, o_l) \in O^2$ corresponds to the probability $e_{\{a_i, b_j\}}(s :: a_i \mapsto o_k; b_j \mapsto o_l)$.

Remark 1.32. To simplify the notation, we will denote the probabilities:

$$e_{\{a_i, \dots, a_k\}}(s :: a_j \mapsto o_j) \equiv e_{(a_i, \dots, a_k)}(o_i, \dots, o_k) \quad (1.47)$$

Sheaf-theoretic contextuality

In the standard contextuality experiments, we are interested in studying the source of the correlations between contexts, i.e. choices of measurements and their observed statistics. In order to isolate the source of potential correlations between the contexts and the outcomes, the standard practice is to limit the overall number of possible sources of such correlations. One type of correlation which can be eliminated in quantum experiments is communication, i.e. the *signalling* between Alice and Bob in the above example. In practice, we can achieve this by spatially isolating these parties.

The consequence of such isolation, or lack of signalling, is that the marginal probability distributions do not depend on the choice of measurements of the other parties. In other words, for any set of inputs U , and any two sets of measurements V, V'

compatible with all elements of U , we should have:

$$e_{U \cup V}|_U(\underline{a}_U) = e_{U \cup V'}|_U(\underline{a}_U) \quad (1.48)$$

for all joint outcomes \underline{a}_U over the measurements of U , where e_W corresponds to the joint probability distribution corresponding with the choices of inputs W for any set W .

Example 1.33. The (2,2,2)-Bell scenario depicted in Table 1.2 indeed satisfies this so-called *no-signalling* condition, since, for instance:

$$e_{(a_1, b_1)}|_{a_1}(0) = e_{(a_1, b_2)}|_{a_1}(0) = \frac{1}{2} \quad (1.49)$$

We then define the notion of (non-)contextuality as follows.

Definition 1.34. A system is said to be *non-contextual* iff there exists a joint probability distribution over X which correctly restricts to all of the e_U 's, i.e., if there exists a global section $e \in P(X)$ such that $e|_U = e$ for all $U \in \mathcal{P}(X)$.

We note that this condition is reminiscent of the definition of a sheaf described in Section 1.1.2 (Definition 1.14). If an empirical model is non-contextual, the global section acts as a hidden-variable model for the observed statistics.

Example 1.35. The example of Fig. 1.2 is *contextual*, i.e. a global probability distribution cannot be defined.

Remark 1.36. This notion of contextuality can be shown to be equivalent to the notion of contextuality defined in terms of non-existence of a homomorphism from a partial Boolean algebras to the Boolean algebra $\mathbf{2}$ [2].

On the no-signalling property

In realistic experiments, the no-signalling condition does not usually hold; this can be due to the unsharpness of the instruments [183] or simply the finiteness of the measurements [183, 52]. As a result, different frameworks have been developed to study contextuality in the presence of signalling. Examples of these are the Contextuality-by-Default framework [52] and the corrected Bell inequalities of the sheaf-theoretic model [183], both of which create a measure of the signalling property of the system. We will describe the Contextuality-by-Default framework in the

subsequent subsection, but first, let's look at way of dealing with signalling in the sheaf-theoretic framework [183].

The intuition is that a signalling system is said to be contextual if the amount of signalling is not enough to make the system “classically explainable”. In sheaf-theoretic terminology, the empirical model is said to be *no-signalling* or *consistent* if every pair of sections in an empirical model satisfies the compatibility condition of (1.48). Given an empirical model e , which is not necessarily compatible, we define the *no-signalling fraction* $\text{NSF} \in [0, 1]$ as the maximal possible value of λ across all of the decompositions of e :

$$e = \lambda \cdot e_{NS} + (1 - \lambda) \cdot e' \quad (1.50)$$

where e_{NS} is a no-signalling empirical model (the multiplication is here point-wise multiplication), and e' can be any empirical model. We then define the *signalling fraction* as:

$$\text{SF} = 1 - \text{NSF} \quad (1.51)$$

The signalling fraction can be seen as the degree of incompatibility of an empirical model, as it measures the departure from a no-signalling, locally compatible model.

Similarly, for any arbitrary empirical model e , we can define the *non-contextual fraction* NCF [183, 4, 12] as the maximal $\lambda \in [0, 1]$ such that:

$$e = \lambda \cdot e_{NC} + (1 - \lambda) \cdot e' \quad (1.52)$$

where, this time, e_{NC} is a non-contextual (and no-signalling) empirical model. In addition, we will also define the *contextual fraction* CF as:

$$\text{CF} = 1 - \text{NCF} \quad (1.53)$$

Then, a possibly signalling empirical model is said to be contextual iff:

$$\text{CF} > \text{SF} \quad (1.54)$$

Remark 1.37. The contextual fraction CF can also quantify a resource from which a quantum advantage can be obtained [3].

Extending to causality

In [130, 79, 80, 5], this formulation of contextuality has been extended to scenarios where structured signalling is allowed, first by allowing sequential operations in [130], then by allowing *definitive causal orders* [79, 5] and even *indefinite causal structures* [79, 80].

Here, we will focus on the case of definite causal order and use the formulation of [79], although the one of [5] is equivalent on the situations of interest in Part II and III^[7]. We start by defining the notion of *party* corresponding to a point in space and time. For example, it could represent a lab, as in the contextuality scenarios, or a sequence of operations done in the same lab.

Each party A will be associated with a set of possible inputs or measurements I_A . The measurements of I_A are assumed to be pairwise incompatible. $X = \prod_A I_A$ will denote the set of all possible measurements. And as in contextuality scenarios, each of the inputs $x \in I_A$ will have an associated set of outcomes O , which we will take to be the same for all possible measurements.

Given a set of parties Ω , we define a *causal order* over Ω as a partial order $\Sigma = (\Omega, \preceq)$ over Ω . This partial order should be interpreted as follows: for any two parties $A, B \in \Omega$, if $A \preceq B$, then the input of A can influence outputs of any measurement chosen by B , but not the other way around. A *causal scenario* is therefore taken to be $(\Sigma = (\Omega, \preceq), X = \prod_{A \in \Omega} I_A, O)$.

Given a set of parties $\omega \subseteq \Omega$, we define its *causal past* as the downward-closed set (with respect to \preceq):

$$\omega_{\downarrow} = \{B \in \Omega \mid \exists A \in \omega. B \preceq A\} \quad (1.55)$$

Then, we define the set of all lower sets Λ_{Σ} as:

$$\Lambda_{\Sigma} = \{\omega_{\downarrow} \mid \omega \in \mathcal{P}(\Omega)\} \quad (1.56)$$

Roughly speaking, each set $\lambda \in \Lambda_{\Sigma}$ corresponds to a set of parties for which a complete history, i.e. sets of inputs and outcomes, can be defined.

We now recall that in contextuality scenarios, for any set of measurements $U \subseteq X$, each measurement $x \in U$ is assumed to be made independently. However, in the case of causal scenarios, the measurements are allowed to depend on the inputs

^[7]Although, the formulation of [5] is applicable to strictly more scenarios than the one of [79].

and outcomes of the preceding events. The approach of [79] is to encode this causal structure within the topology of the base space of the presheaf. Given a causal order Σ , we will then define the topological space \mathcal{L}_Σ as having open sets abstractly defined as:

$$\underline{U} \in \mathcal{L}_\Sigma = (\lambda \in \Lambda_\Sigma, (U_A \subseteq I_A)_{A \in \lambda}) \quad (1.57)$$

where we also require that $U_A \neq \emptyset$ for all $A \in \lambda$.

The idea is that each of these sets \underline{U} gives rise to a well-defined sub-scenario $((\lambda, \preceq), \Pi_{A \in \lambda} U_A, O)$ of the full causal scenario. The condition that $U_A \neq \emptyset$ for all A states that every party in the sub-scenario can “do something” so that every party in its future can use their local experiment.

In addition, we can order these sub-scenarios as follows:

$$\begin{aligned} \underline{U} = (\lambda_U \in \Lambda_\Sigma, (U_A \subseteq I_A)_{A \in \lambda_U}) \subseteq \underline{V} = (\lambda_V \in \Lambda_\Sigma, (V_A \subseteq I_A)_{A \in \lambda_V}) \\ \iff \lambda_U \subseteq \lambda_V \quad \wedge \quad \forall \omega \in \lambda_U. U_\omega \subseteq V_\omega \end{aligned} \quad (1.58)$$

This means that if $\underline{U} \subseteq \underline{V}$, then everything that can happen in \underline{U} is also possible in \underline{V} . We can moreover define the *union* and *intersection* of sub-scenarios \underline{U} and \underline{V} as follows:

$$\underline{U} \cap \underline{V} = (\lambda = \{A \in \lambda_U \cap \lambda_V \mid U_A \cap V_A \neq \emptyset\}, (U_A \cap V_A)_{A \in \lambda}) \quad (1.59)$$

$$\underline{U} \cup \underline{V} = (\lambda_U \cup \lambda_V, (U_A \cup V_A)_{A \in \lambda_U \cup \lambda_V}) \quad (1.60)$$

Remark 1.38. These definitions are directly taken from [79]. However, although they are well-motivated from a physical point of view, the intersection defined in (1.59) is *not always defined*. For instance, let us look at the causality scenario (Σ, X, O) where $\Sigma = (\{A, B, C\}, \preceq)$ is the total order:

$$A \preceq B \preceq C \quad (1.61)$$

and where:

$$I_A = I_B = I_C = \{0, 1\} \quad (1.62)$$

Then, taking:

$$\underline{U} = (\{A, B, C\}, (I_A, \{0\}, I_C)) \quad (1.63)$$

$$\underline{V} = (\{A, B, C\}, (I_A, \{1\}, I_C)) \quad (1.64)$$

Then:

$$\{A \in \lambda_U \cap \lambda_V \mid U_A \cap V_A \neq \emptyset\} = \{A, C\} \notin \Lambda_\Sigma \quad (1.65)$$

which is not a lower set, so $\underline{U} \cap \underline{V} \notin \mathcal{L}_\Sigma$. However, this issue is not easily fixable, and we will leave the task of formulating a better framework as future work.

It is then claimed that \mathcal{L}_Σ forms a locale [79, Proposition 5] (and hence a topological space)^[8].

We say that a function $s : \prod_{A \in \lambda} U_A \rightarrow O^{|\lambda|}$ over a lower set λ respects the causal order Σ iff for all $(i_A)_{A \in \lambda}, (i'_A)_{A \in \lambda} \in \prod_{A \in \lambda} U_A$:

$$(i_A)_{A \in \lambda} \Big|_{B_\downarrow} = (i'_A)_{A \in \lambda} \Big|_{B_\downarrow} \implies s((i_A)_{A \in \lambda}) \Big|_{\{B\}} = s((i'_A)_{A \in \lambda}) \Big|_{\{B\}} \quad (1.66)$$

where we write the (strict) past of B as $B_\downarrow \equiv \{B\}_\downarrow - \{B\}$. This condition states that the past of B is unchanged by its future.

Example 1.39. Let's consider the simple causal scenario $\Sigma = (\{A, B\}, \preceq)$ where the only non-trivial causal relation is $A \preceq B$, and the causal scenario:

$$(\Sigma, \{(A, I_A = \{a_1, a_2\}), (B, I_B = \{b_1, b_2\})\}, O = \{0, 1\})$$

In addition, let's consider the open subset $\underline{U} = (\{A, B\} (I_A, I_B))$. Then, the function $s : I_A \times I_B \rightarrow O^2 \in \mathcal{E}_\Sigma$ defined as:

s	Outcomes
(a_1, b_1)	$(0, 0)$
(a_1, b_2)	$(0, 1)$
(a_2, b_1)	$(1, 1)$
(a_2, b_2)	$(1, 0)$

^[8]Note that this does not follow from the definitions of [79] as, from the above remark, \mathcal{L}_Σ does not define a meet. This could still lead to a locale, but with a different choice of meet.

is a causal function with respect to Σ , since:

$$s(a_1, b_1)|_{\{A\}} = s(a_1, b_2)|_{\{A\}} = 0 \quad (1.67)$$

$$s(a_2, b_1)|_{\{A\}} = s(a_2, b_2)|_{\{A\}} = 1 \quad (1.68)$$

We then define the (pre-)sheaf of causal events $\mathcal{E}_\Sigma : \mathcal{L}_\Sigma^{op} \rightarrow \mathbf{Sets}$ as:

$$\begin{aligned} \mathcal{E}_\Sigma : \quad \mathcal{L}_\Sigma^{op} &\rightarrow \mathbf{Sets} \\ (\lambda, (U_A)_{A \in \lambda}) &\mapsto \{s \mid s \text{ respects the causal order } \Sigma\} \\ \underline{U} \subseteq \underline{V} &\mapsto (s :: (i_A)_A \mapsto (o_A)_A) \mapsto (s|_{\underline{U}} :: (i_A)_A \mapsto (o_A)_A) \end{aligned} \quad (1.69)$$

Each section s of $\mathcal{E}_\Sigma(\underline{U})$ therefore corresponds to a set of consistent histories over the sub-scenario associated with \underline{U} . In this case, the consistency condition expresses the consistency with respect to the causal order Σ .

As in the contextuality case, we then want to consider a probabilistic mixture of possible histories, and hence consider sections of the presheaf $\mathcal{D}_{\mathbb{R}_+} \mathcal{E}_\Sigma : \mathcal{L}_\Sigma^{op} \rightarrow \mathbf{Sets}$. Similarly, we will define a *causal empirical model* as a family of sections $e = \{e_{\underline{U}} \mid \underline{U} \in \mathcal{M}\}$, where \mathcal{M} is a cover of \mathcal{L}_Σ , i.e. $\bigcup_{\underline{U} \in \mathcal{M}} \underline{U} = (\Omega, (I_A)_{A \in \Omega})$.

A standard choice of cover is the following:

$$\mathcal{M}_{local} = \{(\lambda, (\{i_A\})_{A \in \lambda}) \mid \lambda \in \Lambda_\Sigma, i_A \in I_A\} \quad (1.70)$$

This cover will record the statistics of observing the outputs at each stage for any choice of inputs. We will, for instance, use this cover in Chapter 5 when looking at the grammatical parsing process. For this cover, an empirical model is said to be causal or consistent with Σ if the restriction of a section to an earlier stage correspond to the choice of section at this earlier stage, i.e.:

$$\underline{U} \subseteq \underline{V} \quad \wedge \quad \underline{U}, \underline{V} \in \mathcal{M}_{local} \quad \implies \quad e_{\underline{V}}|_{\underline{U}} = e_{\underline{U}} \quad (1.71)$$

Another choice of cover, which we will adopt in Section 3.4, is the following:

$$\mathcal{M}_{global} = \{(\Omega, (\{i_A\})_{A \in \Omega}) \mid i_A \in I_A\} \quad (1.72)$$

In these empirical models, we can only access the final probability distributions

given a global choice of inputs.

Example 1.40. Let's consider once again a causal scenario defined over the causal order $(\{A, B\}, \preceq)$, where the only non-trivial causal relation is $A \preceq B$, and where $I_A = \{a_1, a_2\}$, $I_B = \{b_1, b_2\}$ and $O = \{0, 1\}$. Let's moreover consider the global cover:

$$\mathcal{M} = \{(\{A, B\}, (\{a_i\}, \{b_j\})) \mid i, j = 1, 2\} \quad (1.73)$$

Then, Table 1.3 depicts an example of an empirical model, where each of the rows corresponds to a probability distribution associated with the global choice of input (a_i, b_j) , and the columns are labelled with respect to the observed outcome. This model can moreover be found to be causal with respect to Σ as (removing the curly brackets around singletons and the index Ω for the sake of clarity):

$$\begin{aligned} e_{(a_1, b_1)}|_{a_1}(0) &= e_{(a_1, b_2)}|_{a_1}(0) = 6/13 \\ e_{(a_2, b_1)}|_{a_2}(0) &= e_{(a_2, b_2)}|_{a_2}(0) = 23/65 \end{aligned}$$

	(0, 0)	(0, 1)	(1, 0)	(1, 1)
(a_1, b_1)	0	6/13	0	7/13
(a_1, b_2)	24/65	6/65	7/13	0
(a_2, b_1)	23/65	0	14/65	28/65
(a_2, b_2)	23/260	69/260	42/65	0

Table 1.3: Example of an empirical model causal with respect to to the causal order $A \preceq B$.

Remark 1.41. The notation can quickly become very complex in causal empirical models. Hence, as done in the previous example, any redundant information will be removed in the subsequent chapters whenever it is clear from the context what each of the quantities refers to.

The causal fraction As for the no-signalling property in contextuality scenarios, a generic empirical model will not necessarily be consistent with a given causal order, notably when the probability distributions are obtained empirically. Hence, we will define the notion of the *causal fraction* CausF_Σ with respect to to a causal

order Σ which will quantify how much of the observed statistics is compatible with the causal order Σ . This fraction will be defined as the maximal $\lambda \in [0, 1]$ such that:

$$e = \lambda \cdot e_{\Sigma} + (1 - \lambda) \cdot e' \quad (1.74)$$

where e_{Σ} is consistent with the causal order Σ .

1.2.3 The Contextuality-by-Default framework

We have previously seen that the no-signalling condition imposed on the probability distributions is often too restrictive in practice. Solutions on the sheaf-theoretic side included allowing a small enough amount of signalling into the system or studying systems with a well-defined causal structure. Here, we will describe an alternative way of doing the former, i.e. taking signalling into account, using the framework of Contextuality-by-Default (CbD).

One of the ideas behind the Contextuality-by-Default approach is to extend the notion of contextuality by allowing *direct influence* of the context on the results of measurements. However, for every system in which changing the context results in a change of probability distribution, there is some *contextual influence*. Therefore, one question is to distinguish what counts as “direct influence”, and what is “truly contextual influence”.

In CbD, non-contextual systems are the ones for which one can find a “global explanation” of the system which maximises the probability that distributions corresponding to the same contents coincide. We refer to this minimal amount of contextual influence allowed by the observed probability distributions as *direct influence*, while *contextual influences* will designate any influence due to the context. A system will be contextual if the direct influences are *not enough* to describe the observed system.

We now introduce the standard formalism of Contextuality-by-Default (CbD) (see also [52] for a more general introduction). In this setting, a *content* is a measurement, or more generally, a question with a known set of answers. The *context* gathers all the conditions under which one or several of these questions are asked.

Formally, we start with the concept of a *probability space* (Ω, Σ, μ) , where Ω is

called the sample space, and will correspond to the set of possible outcomes (e.g. set of possible answers to a question), Σ is a σ -algebra over Ω (i.e. set of subsets closed under complementation, countable unions and countable intersections), which we will usually take to be $\Sigma = \mathcal{P}(\Omega)$, and $\mu : \Sigma \rightarrow \mathbb{R}_+$ is a probability distribution. Now, the sample space Ω consists of an abstract collection of objects from which we cannot, for example, calculate expectation value. We then define a *random variable*^[9] over a probability space (Ω, Σ, μ) as a (measurable) function $X : \Omega \rightarrow \mathbb{R}$, where $X(\omega)$ can be seen as the (real-)value of the outcome ω . Then, for any $v \in \mathbb{R}$, we define the probability:

$$P[X = v] = \mu(\{\omega \in \Omega \mid X(\omega) = v\}) \quad (1.75)$$

Similarly, for any $I \subseteq \mathbb{R}$, we define:

$$P[X \in I] = \mu(\{\omega \in \Omega \mid X(\omega) \in I\}) \quad (1.76)$$

Every content q_i in a context c^j gives rise to a random variable R_i^j that takes values from the possible answers to q_i and gives the probability of each answer in the context c^j . So, to make the parallel with the sheaf-theoretic framework introduced in the previous section, for a given measurement scenario we would have:

$$P[R_i^j = v] = e_{c^j|_{q_i}}(v) \quad (1.77)$$

All random variables in a given context are jointly distributed, i.e. they are defined over the same probability space. However, random variables from different contexts are not: they are *stochastically unrelated*. This is the main difference with the sheaf-theoretic framework of contextuality, since here, it does not make sense to question the equality or inequality of marginal probability distributions arising from different contexts, since they are not defined over the same probability space. To talk about random variables that are not jointly distributed, we introduce the concept of *probabilistic coupling*.

Definition 1.42. A *probabilistic coupling* of random variables X_1, \dots, X_n is a set of random variables Y_1, \dots, Y_n which are jointly distributed, and for which the probability distribution of each Y_i agrees with the probability distributions of X_i .

Example 1.43. Here is an example taken from [52]. Consider two unrelated random

^[9]Here we only consider real-valued random variables

variables X_1 and X_2 taking values in $\{1, 2, 3\}$ and $\{1, 2\}$ respectively with the probability distributions are given by:

	$X_1 = 1$	$X_1 = 2$	$X_1 = 3$
P	0.3	0.3	0.4

and :

	$X_2 = 1$	$X_2 = 2$
P	0.7	0.3

Then, we could create a probabilistic coupling Y_1, Y_2 such that the joint probability distribution of Y_1 and Y_2 is given by:

P	$Y_1 = 1$	$Y_1 = 2$	$Y_1 = 3$
$Y_2 = 1$	0.3	0.2	0.2
$Y_2 = 2$	0	0.1	0.2

It can be checked that the marginals of the above joint probability distribution do indeed reduce to the probability distributions of X_1 and X_2 .

Then, given a set of random variables R_i^j , a probabilistic coupling over them will correspond to a hidden variable model of the observed statistics. Note, however, that it is not an analogue of a global section in the sheaf-theoretic framework of contextuality, since the probabilities do not only depend on the content (observable) but also the context it is measured in.

Now, given any set of random variables R_i^j , it is always possible to define (infinitely many) couplings S_i^j [52]. Hence, instead of requiring the existence of a coupling compatible with the observed distributions, we require a “classical-like system” to be a coupling that satisfies certain properties.

In particular, let’s consider a probabilistic coupling S_i^j associated with the observed statistics of a system recorded in R_i^j . Then, since the S_i^j are jointly distributed, the following probability is well-defined for any fixed content q_i and pairs of contexts $c^j, c^{j'}$:

$$P \left[S_i^j = S_i^{j'} \right] = \sum_{v \in V} P \left[S_i^j = v, S_i^{j'} = v \right] \quad (1.78)$$

where V is the set of values the content q_i can take. It can be shown that the above probability is bounded above for any choice of coupling, as:

$$P \left[S_i^j = S_i^{j'} \right] \leq \sum_{v \in V} \min \left(P \left[R_i^j = v \right], P \left[R_i^{j'} = v \right] \right) \quad (1.79)$$

This inequality is, in fact, saturated, i.e. for any pair of random variables R_i^j and $R_i^{j'}$, there exists a coupling $\{S_i^j\}_{c^j, q_i}$ such that:

$$P[S_i^j = S_i^{j'}] = \sum_{v \in V} \min(P[R_i^j = v], P[R_i^{j'} = v]) \quad (1.80)$$

A system $\{R_i^j\}_{c^j, q_i}$ is then said to be *contextual* in the CbD framework iff there exists a coupling $\{S_i^j\}_{c^j, q_i}$ such that for any pair of random variables $R_i^j, R_i^{j'}$, (1.80) is satisfied. If a system is said to be *consistently connected*, i.e. if:

$$P[R_i^j = v] = P[R_i^{j'} = v] \quad (1.81)$$

for any pair of variables $R_i^j, R_i^{j'}$. In most widely studied scenarios^[10], this notion of contextuality collapses to the standard definition of contextuality in no-signalling systems [52, 53].

For a generic system, it is computationally hard to prove the existence or the non-existence of such a coupling, as it requires solving many linear inequalities. We will now focus on a specific type of context-content system, namely *cyclic systems*, for which contextuality can be checked more easily.

In a cyclic system, each context has exactly two contents, and every content is exactly in 2 contexts. The number of contents (or equivalently, the number of contexts) is the rank n of the system. Moreover, again following normal practice in CbD, we will assume that all random variables take values in $\{\pm 1\}$ ^[11].

A cyclic system is known to be contextual in CbD iff [109, 52]:

$$s_{\text{odd}} \left(\left\{ \langle R_{i_j}^j R_{i'_j}^j \rangle \right\}_{j=1, \dots, n} \right) > n - 2 + \Delta \quad (1.82)$$

where $i_j \neq i'_j$ for all j and when $R_{i_j}^j, R_{i'_j}^j$ are well-defined for all j . The s_{odd} function and the quantity Δ are defined below.

^[10]See Remark 1.45 for a discussion about the scenarios in which consistent connectedness and no-signalling are the same notion.

^[11]Every general system can be rewritten as a system with binary variables only [52]; however, in the general case, by making such transformation on a system, it will cease to be cyclic, and the following inequality will no longer apply. There are, however, ways to study the contextuality of such a system [52].

- $s_{odd} : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined as:

$$s_{odd}(\underline{x}) = \max_{\substack{\underline{\sigma} \in \{\pm 1\}^n \\ \mathfrak{p}(\underline{\sigma}) = -1}} \underline{\sigma} \cdot \underline{x} \quad (1.83)$$

where both $\underline{\sigma}$ and \underline{x} are n -dimensional (real) vectors and where $\mathfrak{p}(\underline{\sigma}) = \prod_{i=1}^n \sigma_i$ (\mathfrak{p} can be seen as the parity function of $\underline{\sigma}$). In other words, s_{odd} returns the maximal sum of all its arguments weighted with ± 1 coefficients under the condition that an odd number of negative coefficients are attributed.

- $\Delta \in \mathbb{R}$ is defined as:

$$\Delta = \sum_{i=1}^n \left| \langle R_i^{j_i} \rangle - \langle R_i^{j'_i} \rangle \right| \quad (1.84)$$

where once again $j_i \neq j'_i \forall i$ and $R_i^{j_i}, R_i^{j'_i}$ should be well-defined. The quantity Δ measures a system's "degree of signalling", and a system is consistently connected iff $\Delta = 0$.

We note that equation (1.82) is a generalisation of the inequalities derived in [13] for no-signalling cyclic systems (although they were both proven independently).

Quantifying contextuality

Similarly to the contextual fraction CF defined in the previous subsection, we can define a quantification of the contextuality from the CbD framework. In fact, several measures have been proposed [110], including the non-contextual measure denoted as NCNT2 for a given set of probability distributions $\{R_i^j\}$, defined as:

$$\text{NCNT2} = \min \left(\Delta - s_{odd} \left(\left\{ \langle R_{i_j}^j R_{i'_j}^j \rangle \right\}_{j=1, \dots, n} \right), m \right) \quad (1.85)$$

In the above equation, the quantity m is defined as:

$$m = \min_j \min \left(\langle R_{i_j}^j R_{i'_j}^j \rangle - 2|p_1^j + p_2^j - 1| + 1, 1 - |p_1^j - p_2^j| - \langle R_{i_j}^j R_{i'_j}^j \rangle \right) \quad (1.86)$$

where p_1^j and p_2^j are shorthands for respectively:

$$p_1^j = P \left[R_{i_j}^j = +1 \right] \quad (1.87)$$

$$p_2^j = P \left[R_{i'_j}^j = +1 \right] \quad (1.88)$$

$$(1.89)$$

The advantage of this measure is that we can compare the contextuality of empirical models which are not contextual as the measure can be positive or negative. A negative NCNT2 implies that the model is CbD-contextual, whereas a positive value implies non-contextuality. This measure will be used in Chapter 3.

Quantifying direct influences

As interesting as it is to have criteria for contextuality, we will see in the following Chapters that the amount of signalling in empirical models will be of interest for studying natural language data. We have introduced the signalling fraction from [183] in Section 1.1.2, as well as the “degree of signalling” Δ defined above, but only for cyclic systems.

To obtain a more generic quantification of direct influence within the CbD framework, we introduce another related framework known as M-contextuality (model-contextuality), first introduced in [99]. This framework was inspired by the causal analysis of contextuality of Cavalcanti [36] and the (classical) theory of causality of Pearl [142].

In [99], the author showed that every system of random variables observed in the different contexts can be expressed as a Bayesian network in the form of Fig. 1.4. Here, we treat the contexts as a single random C , and the contents are each modelled by a random variable F_q which is *deterministically* determined by the context variable C and some other latent variable Λ , which corresponds to background knowledge of the system. Such a Bayesian network is called a *canonical model* in [99]. Note that it is important that the latent variable Λ is independent of the context variable C (otherwise, any part of the variable Λ correlated with C can be without loss of generality encompassed by C).

Now, given a canonical model successfully describing a set of observed probabilities, we quantify the *direct influence* of the context variable C on a given content

q as:

$$\Delta_{c,c'}(F_q) = P[\Lambda \in \{\lambda | F_q(\lambda, c) \neq F_q(\lambda, c')\}] \quad (1.90)$$

In turn, a system is said to be *M-contextual* if these direct influences cannot attain their respective minima in a single canonical model compatible with the empirical model. The main result of [99] was to show that this notion of contextuality is, in fact, equivalent to the CbD definition of contextuality.

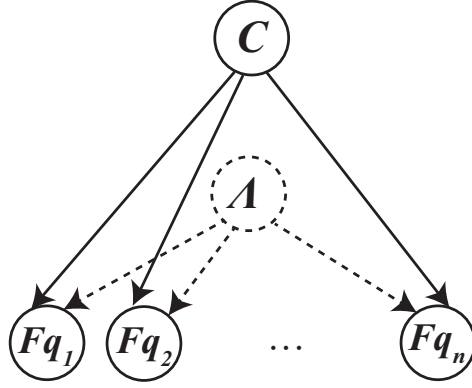


Figure 1.4: Bayesian Network representation of a canonical causal model.

M-contextuality and degree of signalling We now have two ways of quantifying the “direct influence” of a system, namely using the “degree of signalling” Δ from CbD or by using the minimum amount of contextual direct influence $\Delta_{c,c'}(F_q)$ allowed for each content q . As it turns out, these quantities are intrinsically related, and the following is true:

Proposition 1.44. *For a cyclic system with binary random variables taking values in $\{\pm 1\}$, we have:*

$$\Delta = 2 \sum_q \Delta_{c_q, c'_q}^*(F_q) \quad (1.91)$$

where Δ_{c_q, c'_q}^* is the minimum direct influence of the contexts c_q, c'_q associated with content q across all canonical models compatible with the observed distributions.

The proof of this proposition can be found in Appendix C.1.

We note that Δ is only defined for a small class of systems, while the RHS of (1.91) applies to more general systems.

Direct influences and the signalling fraction The notion of direct influence stems from similar motivations as the signalling fraction SF defined in Section 1.1.2, namely, what is the minimal “part” of the observed statistics which can be explained by a no-signalling system. We here show that the signalling fraction gives an upper bound of the degree of signalling Δ in some particular circumstances (including in cyclic systems) and that in the general case, the signalling fraction gives us an upper bound for all of the degree of direct influences in a system^[12]. First, we start with an important remark.

Remark 1.45 (No-signalling and consistent-connectedness). Somewhat surprisingly, although the notion of consistent-connectedness, as defined in [52], is claimed to be equivalent to no-signalling, this is not generally the case. Indeed, a system is said to be consistently-connected iff, for any content X in contexts C, C' , the marginals over X of the probability distributions associated with C and C' are the same. However, in general, the notion of no-signalling is stated as, for any subset $\{X_i\}_{i \in I}$ of contents such that $X_i \in C$ and $X_i \in C'$ for all $i \in I$, then the marginal distribution restricted to all of the X_i coincides. Hence, this distinction only applies if there exist contexts C, C' such that $|C \cap C'| > 1$.

Here is an example of an empirical model which is consistently-connected but signalling.

		(0,0)	(0,1)	(1,0)	(1,1)
a	b	1/2	0	0	1/2
a	b	0	1/2	1/2	0

Having cleared up this distinction, we then state the following results.

Proposition 1.46 (Signalling fraction and degrees of direct influences). *Given statistics of a system for the contexts $\{C_i \subseteq X\}_{i \in I}$ for individual measurements (i.e. contents) X , we have:*

a. *For any system, we have:*

$$\max_{x \in X} \Delta_{C, C'}^*(x) \leq \text{SF} \quad (1.92)$$

b. *If the choices of contexts satisfies $|C_i \cap C_j| \leq 1$ for all $i, j \in I$ and $i \neq j$, then:*

$$\max_{x \in X} \Delta_{C, C'}^*(x) = \text{SF} \quad (1.93)$$

^[12]The results represented here are original at the time of submission of the thesis.

The proof of these claims can be found in Appendix C.2. Moreover, since we already had a relationship between the degrees of direct influence $\Delta_{C,C'}^*(x)$ and the overall degree of signalling Δ (Proposition 1.44), the next results follows.

Corollary 1.47. *In a cyclic system of rank n , we have:*

$$\Delta \leq 2 \text{ SF} \quad (1.94)$$

From Contextuality-by-Default to sheaf-theoretic contextuality

In [50], the author proposed a way of describing signalling empirical models in terms of no-signalling ones within the sheaf-theoretic framework, such that the notion of contextuality in these generated empirical models is equivalent to the notion of contextuality within the Contextality-by-Default framework. This mechanism for creating no-signalling models was coined as *consistentification*. We here briefly describe this procedure.

Recall that in CbD, a cyclic system is contextual is non-contextual whenever it is possible to impose a global probability distribution on the system such that the probabilities $P[S_q^i = S_q^{i'}]$ are simultaneously maximised. This condition can expressed as the possibility of imposing a joint probability distribution on pairs of variables of different contexts that share a content, such that $P[S_q^i = v, S_q^{i'} = v]$ are minimal for every outcome v , and for which marginals coincides with the marginals of the observed variables.

Hence, the process of consistentification consists of creating a new system for which both the contexts and contents of the original system are measurement contexts. The set of observable X is therefore defined as $X = \{(q_i, c^j) | q_i \in c^j\}$, and CbD-contexts and CbD-content correspond to the following set of measurement contexts:

$$\mathcal{M}_c = \{ \{ (q_i, c^j) \in X \} | c^j \text{ is a CbD-context} \} \quad (1.95)$$

$$\mathcal{M}_q = \{ \{ (q_i, c^j) \in X \} | q_i \text{ is a CbD-content} \} \quad (1.96)$$

The probability distributions over the measurement contexts of \mathcal{M}_c are defined as before, i.e. correspond to observed probability distributions. On the other hand, the probability distributions over the measurement contexts of \mathcal{M}_q will be obtained by imposing minimal direct influences on each of the individual contents. This correspondance is illustrated in Fig 1.5. By definition of the S_q^i from above, this system is

no-signalling, i.e., consistently connected.

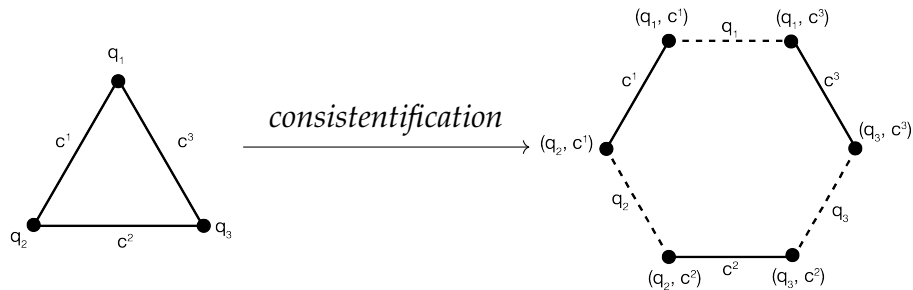


Figure 1.5: Correspondance between the original measurement scenario (left), and the consistentified one (right). On the latter, the solid measurement contexts are the ones inherited from the left-hand measurement scenario, whilst the dashed ones are the ones created from the minimal direct influence condition.

Moreover, the criterion of CbD-contextuality in the original system is, by design, the same as the sheaf-theoretic criterion of contextuality in the generated one.

1.3 Quantum Mechanics as a Process Theory

The goal of this section is to motivate and introduce the notation of Chapter 4. Unlike the previous section, we will now assume the standard Hilbert space formalism of quantum mechanics and provide a categorical description of quantum states and operations.

1.3.1 Features of quantum processes

In Section 1.1.3, we have seen that monoidal categories are very useful in describing processes that can be composed sequentially (modelled using the sequential composition of morphisms) and in parallel (using the monoidal product \otimes). In particular, in the case of quantum processes, we will mainly focus on the category of Hilbert spaces \mathbf{Hilb} . Moreover, in the case of quantum computing, we can restrict ourselves to the category of finite-dimensional Hilbert spaces \mathbf{FdHilb} , which also has some additional “nice” properties, such as the existence of orthonormal bases for all objects of \mathbf{FdHilb} .

on $\mathbb{C}^2 \otimes \mathcal{H}$ as follows:

$$\bullet \text{---} \boxed{U} = \begin{pmatrix} 1 & 0 & \underline{0} \\ 0 & 1 & \underline{0} \\ \underline{0} & \underline{0} & \mathbf{U} \end{pmatrix} \quad (1.99)$$

The intuition is that if the control qubit is in the state $|0\rangle$, the identity on \mathcal{H} is applied, whereas if the control qubit is in the state $|1\rangle$, the unitary U is applied to the target space \mathcal{H} .

Using the monoidal structure, we can also compose morphisms in parallel using the monoidal product \otimes . If two operations U and V are composed in parallel as $U \otimes V$, then it is understood that they are done independently. Now, we have already seen some special (non-unitary) morphisms $|\psi\rangle : I \rightarrow \mathcal{H}$ which represent quantum states. Two states $|\psi\rangle : I \rightarrow \mathcal{H}_1$ and $|\phi\rangle : I \rightarrow \mathcal{H}_2$ can also be composed in parallel as $|\psi\rangle \otimes |\phi\rangle : I \otimes I \rightarrow \mathcal{H}_1 \otimes \mathcal{H}_2$. As with processes, these states are considered independent and known as *product states*. However, given the compound system $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$, not all of the states in \mathcal{H} will be product states. For instance, taking $\mathcal{H}_1 = \mathcal{H}_2 = \mathbb{C}^2$, the Bell state $|\Psi\rangle = \frac{1}{\sqrt{2}} (|0\rangle \otimes |0\rangle + |1\rangle \otimes |1\rangle)$ cannot be decomposed as $|\Psi\rangle = |\psi\rangle \otimes |\phi\rangle$. States that cannot be decomposed as a product state are called *entangled states*.

In addition, we have seen in Section 1.1.3 that the category **FdHilb** also satisfies some extra properties. For instance **FdHilb** is a symmetric monoidal category, i.e. for any two $\mathcal{H}_A, \mathcal{H}_B \in \text{ob}(\mathbf{FdHilb})$, we have $\mathcal{H}_A \otimes \mathcal{H}_B \cong \mathcal{H}_B \otimes \mathcal{H}_A$. The interpretation of this property in terms of quantum systems is quite natural, namely that swapping quantum systems leads to an “essentially equivalent” system in the sense that both swapped and unswapped compound systems share the same physical properties.

Furthermore, the category **FdHilb** is compact-closed, and the duality structure proved very useful. In particular, it is widely known in the quantum mechanics literature that the set of operations $U : \mathcal{H}_A \rightarrow \mathcal{H}_B$ is isomorphic to a set of states $|\Psi\rangle : I \rightarrow \mathcal{H}_A \otimes \mathcal{H}_B$; this correspondence is known as the *Choi-Jamiołkowski isomorphism*^[13]. This equivalence can easily be seen in terms of string diagrams, using the units of

^[13]Here, the isomorphism is understood at the level of sets, i.e. there exists a bijection between the set of quantum states, and the set of quantum operations.

the duality as:

$$\begin{array}{c} \mathcal{H}_A \\ | \\ \boxed{U} \\ | \\ \mathcal{H}_B \end{array} \approx \begin{array}{c} \mathcal{H}_A \quad \mathcal{H}_B \\ | \quad | \\ \boxed{U} \end{array} \quad (1.100)$$

If U is a unitary map, the obtained state under this isomorphism is a *maximally entangled state*, as they can maximally violate the Bell inequalities introduced in Section 1.2.

Example 1.49 (Bell states). In two-qubit systems, certain maximally entangled states are important in quantum protocols such as *quantum teleportation*. These are known as *Bell states* and are defined as:

$$\begin{array}{cc} |\Phi_+\rangle = \begin{array}{c} \text{---} \\ | \\ \text{---} \end{array} & |\Phi_-\rangle = \begin{array}{c} \text{---} \\ | \\ \boxed{Z} \\ | \\ \text{---} \end{array} \\ |\Psi_+\rangle = \begin{array}{c} \text{---} \\ | \\ \boxed{X} \\ | \\ \text{---} \end{array} & |\Psi_-\rangle = \begin{array}{c} \text{---} \\ | \\ \boxed{Y} \\ | \\ \text{---} \end{array} \end{array}$$

(up to normalisation and global phase factors).

1.3.2 Mixed states and density matrices

So far, we have only considered pure quantum states subject to unitary transformation. In realistic systems, however, the quantum states will sometimes interact with their environment in a manner that is not always known or controlled. To deal with this situation, we use *density matrices* and *quantum channels* instead of pure states and unitaries. These are obtained by “forgetting” about the subsystem corresponding to the environment. This construction leads to quantum states which are a probabilistic mixture of pure states. We will here describe the categorical way of defining these states and operations.

Before looking at density matrices, we introduce the categorical notion of *trace* of a morphism.

Definition 1.50. In a compact closed category \mathcal{C} , the *trace* of a morphism $f : A \rightarrow A$,

out, i.e.:

$$\rho = \left(\begin{array}{c} \mathcal{H}_E | \mathcal{H} \\ \Psi^\dagger \\ \Psi \\ \mathcal{H}_E | \mathcal{H} \end{array} \right) \xrightarrow{\cong} \begin{array}{c} \Psi^* \\ \mathcal{H} | \mathcal{H}_E \\ \Psi \\ \mathcal{H}_E | \mathcal{H} \end{array} \quad (1.105)$$

Any state of this form will be known as a *density matrix*. In particular, the analogue of the normalisation condition of (1.97) becomes:

$$\text{Tr}(\rho) = \left(\begin{array}{c} \mathcal{H}_E | \mathcal{H} \\ \Psi^\dagger \\ \Psi \\ \mathcal{H}_E | \mathcal{H} \end{array} \right) = 1 \quad (1.106)$$

As for operations on pure states, we will also restrict permissible operations on density matrices. In particular, we would want these operations to send density matrices to density matrices. Moreover, if an operation is only applied to a subsystem of a larger quantum state, we would also like the resulting state to remain a density matrix. The operators satisfying these conditions are known as *completely positive operators*. From a well-known theorem known as *Stinespring dilation theorem* [171], every completely positive map $V : \mathcal{H}_A \otimes \mathcal{H}_A \rightarrow \mathcal{H}_B \otimes \mathcal{H}_B$ can be decomposed as:

$$\begin{array}{c} \mathcal{H}_A | \quad \mathcal{K} | \mathcal{H}_A \\ \boxed{U^*} \quad \boxed{U} \\ \mathcal{H}_B | \quad \mathcal{H}_B | \end{array} \quad (1.107)$$

where U is a unitary transformation, and \mathcal{K} is a Hilbert space. In addition, in order to preserve the normalisation condition of (1.106), we will say that a *quantum channel* is a complete positive map which is *trace preserving*.

1.3.3 Causality in quantum processes

So far, we have not described any notion of temporal order, as the roles of inputs and outputs in any diagram can be reversed employing the duality and dagger structures. We here want to further restrict the set of physical diagrams by imposing a principle of causality. It turns out that this can be done using the discarding process described above.

We start with a relatively simple intuition. If discarding the entire output of a process would correspond to physically ignoring the outcome of a process, this should be equivalent to ignoring the process in question [104, 105]. In terms of string diagrams, this translates as:

$$\begin{array}{c} |A \\ \boxed{f} \\ |B \\ \underline{\underline{\quad}} \end{array} = \begin{array}{c} |A \\ \underline{\underline{\quad}} \end{array} \quad (1.110)$$

Processes that satisfy this property will be called *causal processes*.

It has been shown that restricting to causal processes is enough to be able to encode the notion of causal relations as described in Section 1.2.1 [104]. First, we will describe the analogues of a party as a pair of input and output systems A_{in}, A_{out} . Then, choosing an input will correspond to selecting an input state in $|\psi_{in}\rangle : I \rightarrow A_{in}$, and similarly, observing an outcome will correspond to $\langle\psi_{out}| : A_{out} \rightarrow I$. Now, given two parties $A = (A_{in}, A_{out})$ and $B = (B_{in}, B_{out})$, the interaction between A and B (regardless of a potential causal order) will be modelled as a causal morphism $f : A_{in} \otimes B_{in} \rightarrow A_{out} \otimes B_{out}$.

Then, we will say that a process f is compatible with $A \preceq B$ iff:

$$\begin{array}{c} A_{in} | \quad | B_{in} \\ \boxed{f} \\ A_{out} | \quad | B_{out} \\ \underline{\underline{\quad}} \end{array} = \begin{array}{c} A_{in} | \quad | B_{in} \\ \boxed{\tilde{f}} \\ A_{out} | \quad | \underline{\underline{\quad}} \end{array} \quad (1.111)$$

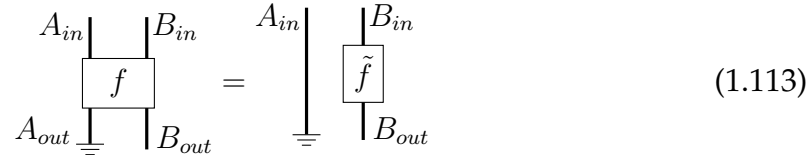
Intuitively, the above equation means that ignoring the subsystem B does not change what happens in the party A . The generic form of such processes is given by [104,

123, 97]:

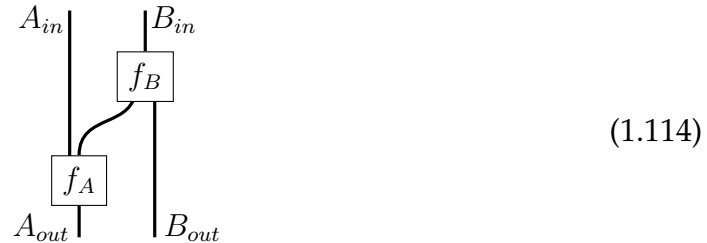


where f_A and f_B are causal processes. When considering the statistics of the measurements of such processes, this notion of causality is indeed equivalent to the notion of compatibility with the causal order $A \preceq B$ as defined in Section 1.2.1 (see Appendix C.3 for more details).

Similarly, we can also say that a process is compatible with $B \preceq A$ iff:



And the generic structure of such processes will be given by:



for causal processes f_A and f_B . These measurement statistics of these processes are also compatible with $B \preceq A$ with respect to the causal relation definition from Section 1.2.1 (see Appendix C.3).

Finally, a process will be said to be no-signalling iff it is compatible with both $A \preceq B$ and $B \preceq A$. The generic structure of no-signalling processes consists of the (monoidal) product of two processes:



We will use these process structures in Chapter 4.

Chapter 2

AMBIGUITIES IN NATURAL LANGUAGES

The English language is ambiguous in several different ways. Examples of the different types of ambiguities are:

- **Lexical ambiguity**

A word is lexically ambiguous whenever we can interpret it in at least two ways. For example, the word *bank* can either mean a financial institution or the side of a body of water in (1).

(1) The bank is far away.

- **Syntactic ambiguity**

A phrase or a sentence is syntactically ambiguous iff it has at least two possible grammatical structures. An example of a syntactically ambiguous sentence is as follows:

(2) She saw a man with binoculars.

In (2), all of the words have definite meanings, but the phrase *with binoculars* can be attached to either *She*, i.e. there is a woman who used binoculars to see a man, or *a man*, i.e. what the woman saw is a man who was using binoculars.

- **Coreference ambiguity**

Texts usually include references to previously mentioned elements. For example, every pronoun such as *he*, *she*, *it* or *they* refers to entities defined from the context. These references can also lead to ambiguous utterances, such as the following sentence:

(3) I put the CD in the computer before it broke.

In (3), the pronoun *it* can equally refer to either *the CD* or *the computer*.

The existence of these ambiguities poses some challenges in NLP, as many of them require knowledge of the world to be disambiguated. For instance, co-reference ambiguities have led to the Winograd Schema challenge, which consists of sentences such as:

(4a) The trophy didn't fit in the suitcase because it was too big.

(4b) The trophy didn't fit in the suitcase because it was too small.

The challenge is then to identify which of the trophy or the suitcase is referred to by the pronoun *it* in each sentence.

In this work, we will focus on studying lexical and syntactic ambiguity. However, similar work has been done regarding co-reference ambiguity (see [120, 121]).

In Section 2.1, we describe how computers and humans process lexically ambiguous words (Sections 2.1.1 and 2.1.2 respectively). We focus on syntactic ambiguities in Section 2.2. In particular, we will introduce the theories of human parsing and the significance of garden-path sentences in Section 2.2.1, and in Section 2.2.2, we look at computational approaches to model human behaviour regarding the parsing of garden-path sentences.

2.1 Lexical ambiguity in linguistics

It is common for words to have several interpretations or multiple entries in a dictionary. When this is the case, the word is *lexically ambiguous*. For example, the word *charge* can be a verb or a noun. As a noun, it has, according to the Oxford Dictionary, the following main possible meanings:

1. A material load; that which can be borne, taken, or received.
2. A load of trouble, expense, responsibility, blame, etc.
3. An impetuous attack

Similarly, as a verb, the word *charge* has the following possible meanings:

1. To load; to cause to bear, hold, or receive.
2. To load heavily; to burden, put anything onerous, troublesome, hateful upon.
3. To attach weight to.
4. To attack impetuously: and senses leading up to it.

Each of these meanings could be further fine-grained, e.g. *charge* in *electrical charge* and in *heavy charge* would both be included in the first definition of the noun *charge* as above, but refer to different things.

Most words commonly used in English are ambiguous, and 99.6% of the words in the British National Corpus [43] are ambiguous. However, this does not create a considerable obstacle for *humans* to understand English. On the other hand, this problem constitutes a significant obstacle for machines in Natural Language Processing.

In this section, we start by reviewing approaches to automatically disambiguate lexically ambiguous words in NLP (Section 2.1.1) and then compare it with the process of human disambiguation as theorised in psycholinguistics (Section 2.1.2).

2.1.1 The challenge of word-sense disambiguation

Word Sense Disambiguation (WSD) is an NLP task that identifies which meaning of an ambiguous word is activated in a given context. WSD was one of the first challenges of NLP as it was crucial in Machine Translation [197]. To see this, consider a word that is ambiguous in the source language but not the target language, then it needs to be disambiguated before one can translate it, e.g. *spring* the season is *printemps* in French, but *spring* the coil is translated as *ressort*. Furthermore, it was shown that lexical disambiguation improves the accuracies of other NLP tasks such as Information Retrieval [203, 31] and Question Answering [154].

WSD can be defined as follows. Given a text T containing the target word w , the aim is to associate w with its intended interpretation, usually taken from a set of definitions or labels of the different possible meanings of w . This task is particularly hard due to the apparent amount of knowledge required and the difficulty of obtaining annotated data.

We will now describe the main approaches in NLP aimed at the WSD task.

Historic approaches

In [161], Schütze divides the task of WSD into two subtasks:

1. **Word-sense discrimination** which aims to classify the contexts in which the intended interpretations are the same;
2. **Word-sense labelling** which aims to label the different classes with a definition.

In [161], the author proposes a completely unsupervised algorithm for solving the problem of word-sense discrimination. The idea behind the approach is similar to the one behind *distributional vectors*.

Indeed, the distributional hypothesis [64, 100, 90] dictates that similar words are found in similar contexts. From there, we can obtain vectorial representations of words, known as distributional vectors, by recording how often a target word w co-occurs with other words in a corpus. These figures correspond to *first-order co-occurrences* [161]. Similarly, we can obtain a representation of a context c by collecting the vector representations of the words in the context c . These vectors correspond to a *second-order co-occurrences*. For first-order occurrences, it has been widely verified that semantically close words are associated with distributional vectors that are close in the vector space [102]. Similarly, we will expect second-order co-occurrence vectors to be close whenever the contexts they represent are semantically close.

The idea of [161] is to identify clusters of context-vectors containing w with different senses. Each sense will then be represented abstractly as the centroid of the associated cluster, i.e. by the average vector of all of the points in the cluster. Then, given a test context c' , we start by creating its context embedding and then select the sense associated with the cluster it is closest to. We, therefore, select the appropriate sense by calculating the distance with all of the centroids. This method resulted in

fairly high accuracies (77.9%) [161], but the obtained clusters do not align with the classifications made by humans, e.g. in dictionaries, and the results are therefore hard to interpret.

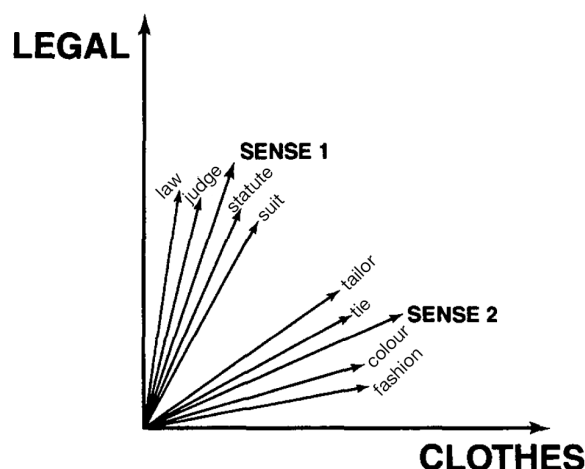


Figure 2.1: Clustering of the contexts for two senses of the word *suit* (adapted from [161]).

Most of the word-sense labelling approaches rely on supervised methods, i.e. dependent on human input or human annotations.

Amongst the most successful supervised settings are *Support Vector Machines* (SVMs) [117, 203, 96], which were first introduced in [30]. SVMs are trained to discriminate positive and negative data points on a vector space by learning the linear hyperplane equation separating positively-labeled and negatively-labelled points.

In the case of the WSD task, we generally want to classify the data points between more than two classes (as a word may have more than two possible meanings). Hence, if a word w has k different senses, the disambiguating task is separated into k binary classification tasks, where the i th task aims to identify whether the intended meaning of w is its i th sense or not. The algorithm gives a confidence score for each sense of w , and the sense with the highest confidence score is then selected.

The dimensions of the vector space correspond to features, e.g. the n -nearest neighbours of w can be encoded in a 5-dimensional vector space. In particular, it has been shown that using features of different nature (e.g. neighbours w , part-of-speech of the neighbours, etc.) is beneficial in WSD tasks [117, 203], see Fig. 2.3 for an example.

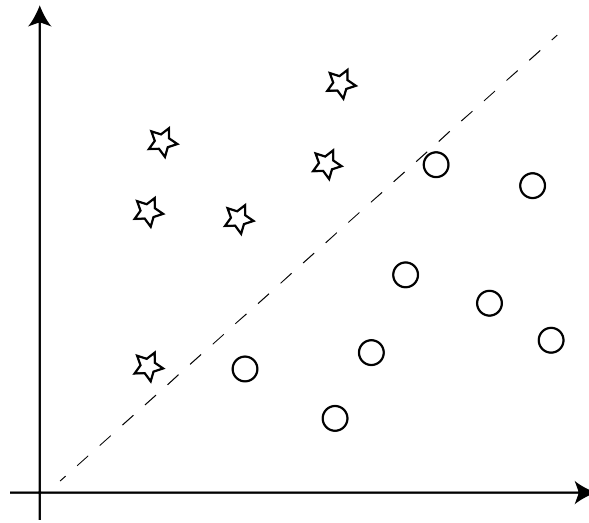


Figure 2.2: Illustration of Support Vector Machines. The two classes of data points (e.g. positive and negative) are depicted in different shapes (e.g. stars and circles). The aim of the SVM is to learn the equation of the dashed line.

The SVMs are then trained using annotated corpora, such as the SemCor corpus [136], a subset of the Brown corpus where each word is annotated by its `WordNet` sense. `WordNet` [134] is a database that contains definitions and examples of the different senses of a word, as well as semantic relations such as hypernymy/hyponymy or word similarity (see Fig. 2.4).

Many WSD systems were evaluated over the `SensEval` benchmark. Four different versions of the `SensEval` tasks have been published, namely `SensEval` [54], `SensEval-2` [55], `SensEval-3` [1] and `SensEval-2007` [9], each consisting of three different tasks:

1. **All-words** in which (almost) all words of a text have to be disambiguated;
2. **Lexical sample** in which only select words have to be disambiguated;
3. **Translation** which is similar to the lexical sample task, but for which, instead of selecting a sense, the WSD system needs to select its translation into a different language, e.g. Japanese.

Each `SensEval` version also provided a target lexicon (i.e. a list of words and their senses), a sense-annotated corpus, and a coarse-graining or fine-graining of senses. Several SVM-based algorithms have been evaluated on the `SensEval` tasks and have

achieved up to 75.2% in SensEval-3 and up to 89.4% in the coarse-grained version of the SensEval-2007 [96].

	Part-of-Speech of neighbours				Surrounding words			
	values: {PRON = 0, V = 1, PREP = 2, ADV = 3}				basis: { <i>account, economy, rate, take</i> }			
	w_{-2}	w_{-1}	w_{+1}	w_{+2}	<i>account</i>	<i>economy</i>	<i>rate</i>	<i>take</i>
Vector	3	1	2	0	0	0	0	1

Figure 2.3: Example of the vector corresponding to the word *interest* in the context *My brother has always taken interest in my work.* (simplified from [203]).

Although they achieve very high accuracies, the supervised approaches are not easily extended to large-scale applications as they require a large amount of manually annotated data. Alternative methods make use of knowledge bases such as (computer-readable) dictionaries, thesauri, or the WordNet database [134].

One of the prominent approaches within this category is the class of *Lesk algorithms* [119, 103, 18, 21]. The idea is that given a target word w in context c , the overlap of the context c with the *gloss* of a sense (i.e. definition and possibly examples) is higher if the sense does correspond to the intended one. Here, the overlap corresponds to the intersections of the set of words in the context and the glosses. For example, given the different glosses of the word *bank*^[1]:

1. a financial institution that accepts deposits and channels the money into lending activities
Examples: he cashed a cheque at the bank, that bank holds the mortgage on my home
2. sloping land (especially the slope beside a body of water)
Examples: they pulled the canoe up on the bank, he sat on the bank of the river and watched the currents

And given the following target context (taken from the BNC):

Cash includes cheque payments, bank transfers and credit card payments .

Then, the algorithm will return the correct intended sense, namely its financial institution meaning; the overlap between the gloss and the target context is underlined

^[1]Example taken from [102]

above. This description is known as the *simplified Lesk algorithm* [103]. In contrast, the original Lesk algorithm of [119] intends to compare the glosses of *all* of the words in the phrase, which increases the algorithm complexity.

As one may expect, only using the overlap of the glosses with the context is a bit crude. Many extensions of this approach have been proposed, for example by also considering the words in glosses of related words (obtained from WordNet) [18] or by using distributional or neural representations of contexts and glosses [21] and taking the cosine of vectors as the measure of overlap.

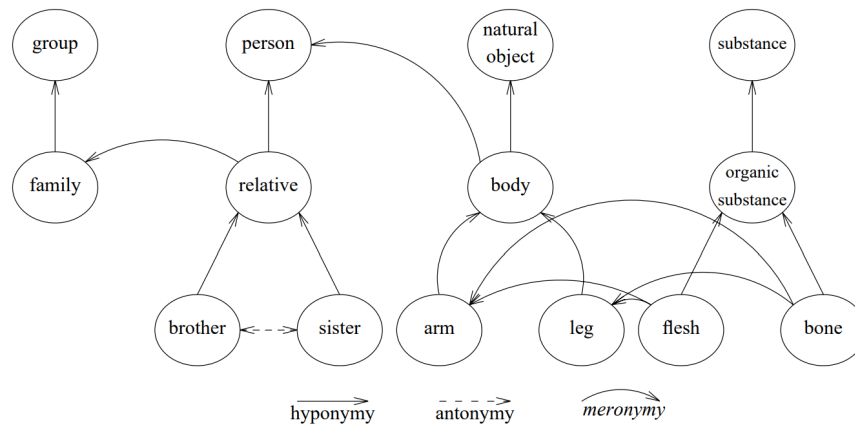


Figure 2.4: Illustration of the WordNet knowledge base. (taken from [135])

Neural approaches

The state-of-the-art approaches in NLP differ from the previously described methods as they vastly rely on artificial *neural networks*. As for distributional approaches and SVMs, the meanings of words and sentences are stored as vectors known as embeddings. The entries of these vectors are learned from input/output pairs from a training set, where the input goes through a network of nodes (or artificial neurons) with tunable activation strength – the details of this process depend on the architecture of the neural network.

One of the first instances of successful neural network architecture is the *Recurrent Neural Network* (RNN), in which a sequence of tokens is input and processed from left to right (see Fig. 2.5a). Hence, at any stage, the network can retain information from all previous words. RNNs were a way to escape the sparsity of n -grams

co-occurrences in corpora [23, 133] and were soon found to give meaningful representations of words. For instance, the `word2vec` vectors were found to predict semantic relations between words accurately [133].

However, in “standard” RNNs, the contribution of a given token to the gradient can either tend to 0 or ∞ as training time increases, which makes them impractical for dealing with large texts. The *Long-Short Term Memory* (LSTM) architecture aims to solve this problem by learning to “forget” information that is no longer relevant [92].

In addition, by restricting the processing of words strictly from left to right, some of the long-distance dependencies may be lost from this forgetting mechanism. For this reason, *bidirectional* LSTMs (biLSTMs) were proposed as an alternative, where two LSTMs, one going from left to right and the other from right to left, are combined.

In [131], the authors proposed an extension of the `word2vec` word-embeddings from [133] by creating embeddings of contexts, referred to as `context2vec` context-embeddings. These context-vectors were then used in WSD as follows. Given a target word w , we can collect all of the context-embeddings associated with each of the occurrences of each sense s of w from a sense-annotated corpus such as SemCor [136]. Similarly, we can create a context-vector for the target context. We then compare this target context-vector to all of the relevant context-vectors obtained from the annotated corpus, and we select the sense associated with the context-vector closest to the target context-vector.

In [145], the authors used a similar algorithm, but using the `ELMo` (Embeddings from Language Models) *contextualised embeddings* (which represents words in contexts) instead of the context-embeddings (which represent the context assuming all words are context-independent). The approach of [145] also differed from the one of [131] as sense-embeddings of a target word w were obtained by averaging the contextualised word-embeddings of occurrences of w in SemCor which corresponded to the same sense.

The above approach has the major flaw that only a 16.11% of `WordNet` senses are found in SemCor [124]. The solution of [145] was always to select the most common sense for each unseen word. To obtain a representation of an unseen `WordNet` sense, the authors of [124] make use of the structure of the `WordNet` lexical database. In-

deed, each of the senses (corresponding to a lemma, its part of speech, and its gloss) is organised in synsets, which include synonymous senses. Each synset has a set of *hypernyms*, e.g. dog_n^1 is a hypernym of pug_n^1 , which is, in turn, part of a larger *lexname*, e.g. dog_n^1 is part of *noun.animal*. We can obtain the representation of each of these abstraction levels by averaging the context-embeddings associated with all of the senses included in them and are present in the annotated corpus. The representations of the missing sense would then be abstracted by the representation of its first hypernym or lexname for which a representation exists.

Furthermore, motivated by the intuition behind the Lesk algorithms, some neural approaches have also used glosses as a knowledge source. It was shown that including the gloss embeddings on top of the context-embeddings using co-attention mechanisms [125, 126] or by simple concatenation [124] improves the performance of WSD algorithms.

In 2017, Vaswani et al. introduced a novel neural network architecture known as the *transformer* [187]. In particular, this new architecture allowed parallelisation of the training process by allowing all-to-all connectivity of the artificial neurons (see Fig. 2.5b). The parallelisation of the training process opened the opportunity to train the neural network using a substantial amount of data. For example, the Google language model BERT (Bidirectional Encoder Representations from Transformers) was trained over 3.3 billion tokens, while the largest version of GPT to date was trained over 449 billion tokens. The *pre-training* process conducted by Google or

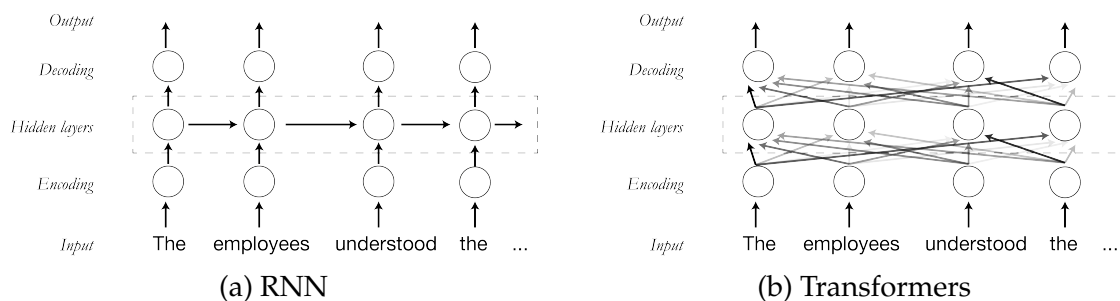


Figure 2.5: Schematics of the differences between recurrent and transformers neural networks architectures.

OpenAI is done by tuning the neural network's attention weights (i.e. the strength of the connection between two nodes) to solve a generic task.

For example, BERT was trained simultaneously on mask prediction and next-sentence prediction tasks. The mask prediction task is as follows: the neural network is presented with a sentence where one or several words^[2] are *masked*, and its goal is to predict the values of these masks. For example, the language model could be presented the input:

```
Paris is the [MASK] of France.
```

and will attempt to predict the word *capital*. In the next sentence prediction task, the language model is shown two sentences S_1 and S_2 and needs to decide whether S_2 follows S_1 . For example, if:

```
S1 =The man went to the store.  
S2 =He bought a gallon of milk.
```

the neural network should output `True`, i.e. S_2 indeed follows S_1 ; but in the case of:

```
S1 =The man went to the store.  
S2 =Penguins are flightless birds.
```

the output should be `False`.

From the pre-trained language models, two main ways exist to solve the WSD problem. First, as for the LSTMs described earlier, given a text input, the transformer neural network will return a contextualised word-embedding. Hence, we can use these embeddings as in the previously described algorithms. For example, the same Nearest Neighbour algorithm described above for LSTMs performed better using BERT embeddings than using `context2vec` [131] or ELMo [145] embeddings [124].

Another possibility is to *fine-tune* the pre-trained models by training the language model for a more specific task, such as WSD, using a much smaller training

^[2]To be more accurate, the masks are tokens and not words; the distinction is not important for the rest of this work.

dataset than the one required for pre-training. In [200], this approach was taken on the language model BERT, where the fine-tuning process was as follows. The training set consisted of tuples (c, d, l) where c is a context containing a target word w , d is a definition of a sense of w , and l is the label in $\{yes, no\}$ corresponding to whether the intended meaning of w in c corresponds to the definition d . The target context c and definition d are then fed into the neural network, which will then be trained to predict the correct labels l . This approach achieved an accuracy of 72.3%, which is comparable to the performance of SVMs, but does not use annotated corpora.

2.1.2 The human disambiguation process

In this section, we review the psycholinguistic theories on the lexical disambiguation process of humans. In particular, we will focus on the differences in the processing of ambiguous nouns and verbs and words with different “levels of ambiguity”. Indeed, two interpretations of an ambiguous word could be completely unrelated, such as *bank* in *bank account* and in *river bank*, or somewhat related, for example, *book* in *interesting book* and in *hardback book*. In the case of *book*, both expressions refer to the same entity, but the former relates to information content, whereas the latter interpretation relates to the properties of the physical object.

When the interpretations of an ambiguous word are unrelated, the word is said to be *homonymous*. In contrast, if its interpretations are related, it is said to be *polysemous*. We will, for the rest of the thesis, adopt the terminology of the literature where *meanings* will refer to unrelated interpretations, *senses* will refer to related interpretations, and *interpretations* can refer to both meanings or senses.

Psycholinguists study the disambiguation process of lexically ambiguous words using eye-tracking data. In such settings, the participants are presented with a text on a screen and asked to read it. The eye-tracker will then record the movements of the participant’s eyes and the lengths (and order) of the eye fixations on different zones of the screen (which usually coincide with each of the words of the text).

The prominent figures of interest will be a target word’s first-pass and second-pass fixation times. The former corresponds to the time spent on a word reading from left to right, and the latter corresponds to any additional fixation on the target region (i.e. when the reader has to go “back” to the target word).

Homonymous nouns

Most studies on lexical ambiguity in psycholinguistics focus on the processing of homonymous nouns, i.e., nouns that have unrelated meanings.

The main effect detected from eye-tracking experiments over homonymous nouns relates to the frequencies at which each meaning occurs. In particular, we observe a slowdown whenever the interpretation that has to be selected is uncommon [69, 47]. For example, in the following (taken from [69]), the sentence in (1a) was read faster than it (1b).

(1a) Playing so loudly, the wedding band upset the groom.

(1b) Looking so tarnished, the wedding band upset the groom.

This suggests that a homonymous noun's most common meaning is activated more easily than uncommon ones. We will emphasise here that this bias in meaning selection is *independent of the context* and only depends on the reader's experience, i.e. how often they have seen each of the different meanings in their lives.

Moreover, a slowdown also occurs whenever the meanings of the target word were *balanced*, i.e. when the meanings are roughly as common as each other, for example *palm* the type of tree or *palm* the part of the hand [156]. In fact, in [156], the authors showed that if the meanings were balanced, the slowdown occurs when the reader encounters the ambiguous word. In contrast, in the case of non-balanced words, a slowdown only occurs in the *disambiguating region* whenever the subordinate (i.e. the less common) meaning has to be activated. This result has also been replicated in [47].

This finding suggests multiple interpretations of a homonymous noun can be activated simultaneously. However, the activation levels will not be the same for all meanings, and a reader will give the most common meaning a higher "rating". If multiple meanings are equally probable, they will compete, and this competition will create difficulty when reading the ambiguous word. This describes a *parallel ranked* process of disambiguation. Moreover, when the context suggests that a subordinate meaning has to be selected, the reader encounters difficulty when seeing the disambiguating context, as they will have to readjust the activation levels. On the other hand, if the intended meaning is the dominant one, i.e. the most common one, no difficulty should occur.

Finally, the position of the disambiguating context also plays a role in the disambiguation process. In particular, if the disambiguating context is *before* the target word, the fixation times are higher on the target word (but low on the disambiguating region) than when the disambiguating region is *after* the target word [69, 47]. However, the reading time of the whole sentence was higher whenever the disambiguating context is *after* the target word [69, 47].

This would suggest that the disambiguation process is incremental, i.e. the reader will readjust the activation weights of each possible meaning as more information is known.

Polysemous nouns and underspecification

We now look at the disambiguation process of polysemous nouns, which, somewhat surprisingly, carries some stark differences from the disambiguation process of homonymous nouns.

In particular, no difference in reading times has been observed between common and uncommon senses [69], nor any difference between concrete and abstract senses [69], well-known senses and senses created through rules [72]. In all of these cases, the reading times observed are comparable with the reading times of unambiguous words [69, 72].

This suggests that, instead of having a parallel-ranked representation of the different senses of a polysemous word, the reader will always start by selecting an *underspecified* interpretation of the word, which includes all of its possible senses. In the theory of underspecification, the context then shapes the salient interpretation of a polysemous word.

Although the interpretation is underspecified within a sentence, if the sentential context of a polysemous word is not enough to select a sense, the possible senses behave like meanings in the following sentence, i.e. relative frequencies start to have an impact on the reading times [71]. This phenomenon is referred to as a *homing-in* stage.

This observation suggests that, even though polysemous words are processed faster, commitment to an interpretation, i.e. selecting the most appropriate sense, is only done at the end of the sentence, whereas we recall that this is meaning selection occurs almost instantly in homonymous nouns.

Remark 2.1. Parallel unranked models of the disambiguation of polysemous words have also been proposed. However, they suffer from several drawbacks. In particular, there is no obvious way to select the most appropriate meaning from them. In addition, parallel unranked models would predict that the more possible senses a word has, the more difficult it should be to read. However, this prediction was not in line with eye-tracking data [71].

Disambiguating verbs

Most of the psycholinguistics literature on lexical ambiguity focuses on ambiguous nouns, and seldom research looked at the disambiguation process of more complex ambiguous types such as verbs and adjectives. In [174], the authors investigated words with multiple possible grammatical types (e.g. *wacth* could be a noun or a verb), and the authors of [139] looked at the process of disambiguating adjectives. Here, we will mostly focus on the disambiguation of ambiguous verbs, which was studied in [146].

In particular, in [146], the authors showed that the processing of ambiguous verbs (both polysemous and homonymous) differed from the processing of ambiguous nouns. Indeed, we observe a general slowdown in reading time of ambiguous verbs, compared to ambiguous nouns, which shows that ambiguous verbs are more complex to disambiguate.

The effect of frequency was much smaller for homonymous verbs than it was for homonymous nouns. Indeed, no significant difference in first-pass reading times of the target verb between common and uncommon meanings was observed. In second-pass reading, only a mild frequency effect was detected, where less backtracking is observed when the intended meaning is the dominant one. In addition, the main difficulty did not occur in the verb region but was slightly delayed until the object of the verb was encountered. This finding suggests that the disambiguation of homonymous verbs is not immediate, contrary to homonymous nouns, but the reader waits until the arguments of the verbs are known. In particular, since meaning selection is delayed, both the dominant and subordinate meanings “have the time” to be activated, which diminishes the frequency effects.

For polysemous verbs, the disambiguation was delayed even further, and the slowdown in first-pass mainly occurred at the end of the phrase or the sentence. In

addition, similarly to polysemous nouns, no significant frequency effect could be observed during the initial analysis. This delay suggests once again that polysemous verbs will initially activate an underspecified meaning, which is then made more and more specific as information from the context emerges. Some minor frequency effects occurred on second-pass, which suggests that the reader will home-in on a chosen sense at the end of the sentence, which consequently behaves more like a homonymous verb on reanalysis.

Many hypotheses can be advanced to explain the relative difficulty of processing verbs compared to nouns. First, this general increase in difficulty is not restricted to ambiguous cases. When reading a sentence, readers will generally spend longer on the main verb of the sentence than any other words [155]. In addition, many studies suggest that the meaning of verbs is highly dependent on their arguments [76], making its interpretation much more variable [76, 75, 74]. For instance, in the case of a mismatch between the arguments of a verb and its standard interpretation, e.g. in (2b) as opposed to (2a), it is the verb that tends to acquire a metaphorical interpretation.

(2a) The mule shivered

(2b) The car shivered

This apparent complexity might be deeply rooted in language acquisition. Indeed, it is well-established that children will learn nouns before verbs [74, 26] and have more difficulty using verbs [178]. This factor could explain the overall difficulty of disambiguating verbs.

2.2 Human parsing & garden-path sentences

In this section, we will describe another type of ambiguity, namely *syntactic ambiguity*. This ambiguity arises whenever several *gramamtical structures* are simultaneously possible. An example of a syntactically ambiguous sentence is:

(1) She saw a man with binoculars

This sentence either means:

(1a) She used binoculars and saw a man

(1b) She saw a man, and he was using binoculars

In addition, syntactic ambiguity does not always occur at the sentence level. Sometimes, a sentence can also be *locally* ambiguous. For example, consider the sentence fragment:

(2) The artist painted [...]

The verb *painted* can either be transitive, i.e. take an object such as *a portrait*, intransitive, i.e. does not take any object, or even be in the passive voice, e.g. the artist is the thing that is painted.

Certain types of sentences, known as *garden-path sentences*, have been used in psychological research to unmask the processes behind grammatical parsing. These are sentences for which humans initially parsed a locally ambiguous fragment incorrectly.

For example, in the following sentence, the phrase *the contract* is initially parsed as the object of the verb *understood* and is eventually found to be the subject of the verb phrase *would change*.

(3) The employees understood the contract would change

In this section, we review the existing literature on garden-path sentences. In Section 2.2.1, we introduce the different human parsing theories and the evidence supporting them. These theories will motivate our models described in Part III. In Section 2.2.2, we describe a popular framework of computational linguistics, namely surprisal theory, which has been applied to predict reading times of garden-path sentences. We give some history and motivation for this framework and its drawbacks.

2.2.1 Psycholinguistics theories of parsing

In [25], Thomas Bever exposed an overview of his theories of *perceptual strategies*, which are mechanisms that allow humans to convert (external) linguistic structures to (internal) perceptual representations of meaning. He believed these perceptual mechanisms are more fundamental than linguistic structures like grammar

and are the first to be acquired by children. Therefore, language acquisition would correspond to learning labels and rules corresponding to these perceptual strategies. Among these strategies, he explains that one of the first steps of language understanding is parsing a sequence of words and sounds (external structures) into groups associated with a fundamental role, such as actor, action, object, or modifier (internal structures). A subsequent rule is to associate a “N...V...(N)” sequence with the “actor-action-object” roles as soon as possible unless markers suggest that the passive voice is used or that the first phrase modifies the main clause. To illustrate his claim, Bever describes a series of linguistic behaviour and experiments testing them. Among these behaviours, he describes the difficulty of parsing sentences such as:

- (1) The horse raced past the barn fell

The difficulty of parsing would be due to non-conformity with respect to those perceptual strategies. These sentences were later referred to as *garden-path sentences*.

Generally speaking, a garden-path sentence is a syntactically unambiguous sentence for which the reader is “led down a garden-path”, i.e. they are forced to adopt the wrong syntactic structure at some initial stage. After Bever’s original work, these sentences have been widely used in psycholinguistics to uncover mechanisms at the heart of human parsing by studying what induces errors for readers [68, 152, 93, 138, 181, 147, 73, 173].

Some specific types of garden-path sentences are widely studied in psycholinguistics. These are exemplified as follows.

- **NP/S sentences**

- (2) The employees understood the contract would change.

Here, the main verb *understood* either takes a noun-phrase (NP), such as *the contract*, as an object, or a sentential (S) object, such as *the contract would change*.

- **NP/Z sentence**

- (3) As the woman read the magazine amused the editors.

In this case, the main verb *read*, either takes an NP as an object, e.g. *the magazine*, or no object at all – this is denoted by (Z) for “zero”.

- **MV/RR sentences**

(4) The soldiers warned about the dangers conducted the raid.

In this case, the underlined verb *warned* is either the main verb (MV) or part of a relative clause (RR). The example (1) is also an MV/RR sentence.

In this thesis, we will mostly focus on sentences of type NP/S and NP/Z.

What has been shown unambiguously in several studies is that these different types of garden-path sentences are read with different levels of difficulty. In particular, NP/S sentences were read faster than NP/Z sentences, so NP/S sentences are more straightforward to parse than NP/Z ones [173, 83]. What is more debated in the literature is why this is the case. One hypothesis is that it is due to the nature of the changes needed to obtain the correct parse [173].

An underlying problem is to identify why a garden-path effect occurs in the first place.

The first thing to mention is that local ambiguity is not the main cause of difficulty. In fact, the presence of syntactic ambiguity can make a sentence *faster to read* than its unambiguous variants [179]. This feature distinguishes syntactic disambiguation from lexical ambiguity resolution (see Section 2.1.2). The intuition is as follows. Let’s assume that a given fragment has two equally likely possible syntactic structures. If the sentence is only locally ambiguous but globally unambiguous, then half of the time, the reader will initially select the “wrong parse”. Hence, reanalysis is triggered half of the time. On the other hand, if the sentence remains globally ambiguous, then no analysis is ever required. The sentence is then read faster on average.

However, several contributing factors to the existence of garden-path effects have been identified, among which lexical [68, 93, 138, 181, 73], plausibility [147, 73] and discourse biases [68, 180]. The main quoted factor is the frequency bias of its main verb [68, 93, 138, 181, 73]. For instance, the verb *hear* would have a higher bias towards taking an NP as an object than the verb *claim*. Hence, sentences with the

main verb *hear* would be more likely to create an NP/S or NP/Z garden-path sentence than sentences with the main verb *claim*. This phenomenon is exemplified in the following, where (5a) does not create considerable difficulty, whereas (5b) does.

(5a) The officer claimed the alarm was a surprise.

(5b) The officer heard the alarm was a surprise.

Similarly, the difficulty of selecting the correct parse in garden-path sentences increases if the object of the NP reading is deemed plausible [147, 73]. For example, (6a) is harder to read than (6b).

(6a) As the woman read the magazine amused the editors.

(6b) As the woman sailed the magazine amused the editors.

With regards to *how* do humans resolve the garden-path effects, two main categories of procedures have been proposed, namely *serial processing* [68, 152] and *parallel processing* [101, 184]. Advocates of the serial strategy argue that, at any given stage, the reader will create a *single parse* which can be completed [68, 152], and that certain conditions can be imposed on the partial structure (e.g. minimal attachment states readers never add unnecessary nodes to the structure under construction).

In the case of parallel processing, when syntactic ambiguity occurs, at least two or more parses are constructed in parallel. However, since we do observe a garden-path effect in sentences such as (1)- (4), this implies that these structures have to be *ranked*. Otherwise, the less likely parse having been constructed already, it should not be hard to extend the “correct” parse. The difficulty then comes from some “reranking mechanism”, weights have to be transferred from the previously likely but incorrect parse to the correct parse.

It is quite difficult to distinguish between probabilistic serial sentence processing (i.e. if two or more parses are possible, they are chosen with their respective degree of likelihood) or ranked-parallel processing (i.e. all of the possible parses are created at the same time, but with different accessibility levels) [77]^[3]. In our model described in Chapter 5, we take an approach compatible with both interpretations.

^[3]This is similar to the fact that for a single quantum state, a superposition of basis state is indistinguishable from a probabilistic mixture of the same basis states if only basis measurements are available.

Finally, there is also the question of how the reader obtains the correct parse. In [83], the authors compare two different hypotheses, namely *repair*, where the reader will amend a given parse to obtain the correct one, or *reanalysis*, where the reader will reparse a sentence fragment in the view of incoming information. The authors of [83] presented evidence supporting the reanalysis hypothesis. The study uses a *locality bias* (i.e. it is easier to attach clauses that are close together), such that, by adding some extra words between the start of clause marker *As* in (3) and the ambiguous NP *the magazine*, the garden-path effect decreased in NP/Z sentences, whereas no such effect occurs in NP/S sentences.

2.2.2 Surprisal predictions for garden-path sentences

Psycholinguistic studies have shown that one of the main factors influencing reading time is *predictability* of a word in a context [56]. Indeed, words are read faster if found in a context that makes them predictable than in contexts where they are not [56]. For example, the word *shark* is read faster in (7a) than in (7b).

(7a) The coast guard had warned that someone had seen a ...

(7b) The zoo keeper explained that the lifespan of a ...

In [168], the relation between predictability and reading time was *logarithmic*. This result was obtained by looking at eye-tracking times of a subset of the Dundee dataset (containing newspapers) and self-paced reading times for subsets of the Brown corpus (containing texts of various genres). The reading times correlated with the conditional probabilities of encountering a word w_{n+1} in the context $c = w_1 \dots w_n$. This then motivates the use of the *surprisal*, defined as:

$$S(w_{n+1}|w_1 \dots w_n) = -\log_2 P[w_{n+1} | w_1 \dots w_n]$$

as a predictor for reading time. The authors from [168] demonstrated that:

$$S(w_{n+1}|w_1 \dots w_n) \propto RT(w_{n+1}|w_1 \dots w_n)$$

Surprisal, also known as self-information, originates in Shannon's theory of information [165]. In this theory, surprisal is defined as the *quantity of information* entailed by knowing the value $X = w$, where X is defined as a random variable

selecting the following word in the context $w_1 \dots w_n$. The intuition is that a very predictable word is not surprising and, therefore, does not carry out a lot of information. On the other hand, if a word is not predictable, it should significantly increase the amount of information available to the reader.

Predictability has historically been estimated from *cloze tasks* [175], where human participants are asked to complete a sentence or a piece of text. However, cloze tasks fail to estimate the predictability of highly improbable (probability $< 5\% - 10\%$) words and constructions [168]. Therefore, data from such tasks are not reliable for studying garden-path sentences.

In [168], the authors instead decided to collect statistics (trigram probabilities with a bigram cache) from text corpora (e.g. BNC) to obtain word predictability. These probabilities, however, only take local context into account and do not necessarily mirror the way humans assign probabilities [168].

With the advent of neural language modelling, computational linguistics soon realised that they could use predictions from language models instead of collecting predictability from cloze tasks.

Hale [87] was the first to suggest using surprisal for predicting the slowdown in garden-path sentences. To do so, the author used the probabilities coming from a probabilistic Earley parser. However, the correlation between the magnitude of surprisal and reading time was not investigated [87].

In [185, 186, 95, 14], the authors studied the empirical correlations between surprisal from language models and self-paced reading times of garden-path sentences. Although surprisal calculated from language models can predict the existence of a garden-path effect (i.e. a higher reading time in the critical region in garden-path sentences compared to their unambiguous version), it consistently underestimates its magnitude [185, 186, 95, 14]. In addition, although predictions are mostly lower for NP/S than for NP/Z [185, 95], no statistical difference has been found between the predicted garden-path effects of NP/S and NP/Z sentences. In fact, in the study presented in [186], the average garden-path effect for NP/S sentences was lower than that for NP/Z sentences.

Many possible reasons for the discrepancies between surprisal and reading times have been advanced:

1. Surprisal is not the main predictor for reading times [186];
2. The human parsing process is not strictly incremental and reanalysis or backtracking is necessary [185, 186, 95, 14];
3. The statistics used to calculate surprisal are not adequate representations of how humans assign predictions [14].

In [186], authors have compared the predictions from surprisal with predictions from alternative incremental measures of dissonance that have been proposed in the past, such as entropy and entropy reduction [88, 89, 67]. The idea is that reading time is also modulated by how much uncertainty is removed by adding an extra word to a sentence fragment [88, 67]; the higher this differential of uncertainty, the higher the reading time. What they found is that surprisal outperformed the entropy-based models by quite a distance, as the entropy-based measures did not predict a garden-path effect at all [186]. This shows that if the disambiguation process is truly incremental, surprisal would be the best-known prediction factor.

One other factor that authors of [186] have explored is whether the predictions from large language models suffer from the same drawbacks as the cloze tasks, namely that low probabilities predictions are effectively not predicted – and therefore, LLM predictions are not reliable for rare or complex constructions. Indeed, no ceiling effect was observed [186], therefore confirming that this was not the main source of error in the surprisal calculations.

Finally, in [14], the influence of the *syntactic surprisal*, as opposed to the lexical surprisal defined above, on the garden-path effect predictions was investigated. To do so, the authors defined a surprisal measure based on the probability of obtaining a given Combinatory Categorical Grammar (CCG) supertag for the last word of a sentence fragment, conditioned on having the usual (lexical) context consisting of the previous words. In other words, the syntactic surprisal is defined from a new probability distribution:

$$P[c_{n+1} | w_1 \dots w_n] = \sum_{w_{n+1}} P[t(w_{n+1}) = c_{n+1} | w_1 \dots w_{n+1}] \times P[w_{n+1} | w_1 \dots w_n] \quad (2.1)$$

where, as before, w 's runs over the vocabulary V , c 's runs over the set of CCG supertags C , and the map $t : V \rightarrow C$ associates a word with a CCG type. In addition, since the CCG supertag of a word is not necessarily unique, all possible supertags

were considered in the syntactic surprisal calculations. Therefore, the syntactic surprisal is overall defined as:

$$S_{synt}(w_{n+1}|w_1 \dots w_n) = -\log_2 \sum_{c_{n+1}} P[c_{n+1} | w_1 \dots w_n] \times P[t(w_{n+1}) = c_{n+1} | w_1 \dots w_{n+1}] \quad (2.2)$$

The syntactic surprisal provided more accurate predictions than the lexical surprisal alone, and combining both lexical and syntactic surprisal improved the predictions [14]. However, even with these improvements, the garden-path effect was still widely underestimated [14], and thus, including syntactic surprisal did not fully resolve the previous issues. In Chapter III, we will give an alternative model based on sheaf-theory which achieves better predictions.

Part II

LEXICAL AMBIGUITY

Chapter 3

ASPECTS OF THE LEXICAL DISAMBIGUATION PROCESS

In this chapter, we want to create a model of the human lexical disambiguation process. Moreover, our goal is to make this model quantum native, so that we can explore the potential of quantum simulations in the next chapter. Here, we study natural language data using mathematical frameworks arising from quantum mechanics (Sections 3.2 to 3.4). By doing so, we create a parallel description of linguistic features in terms of quantum ones.

Moreover, we recall from Section 2.1.2 that humans do not disambiguate words of different grammatical types, or different levels of ambiguity in the same way. However, the mainstream NLP approaches to word-sense disambiguation introduced in Section 2.1.1, disambiguate all words in the same way, regardless of their part-of-speech, and whether the interpretations are related or unrelated. With our approach, we study the linguistic data for words of different parts-of-speech (i.e. nouns and verbs) and different levels of ambiguity (i.e. polysemous and homonymous words). We then compare our results with the theories of psycholinguistics presented in Section 2.1.2.

More specifically, we start by looking at the potential *contextuality* of lexical ambiguity data (Section 3.2). In Section 3.3, we observe that the *signalling* property of

the collected data is a quantity of interest and study it further. In Section 3.4, this analysis is extended by studying the structure of the observed signalling, via its *causal structure*.

3.1 Methodology

In all of the different studies carried out in this chapter, we will use a common interpretation of the analogue of quantum measurements. The idea is to take parties to represent *grammatical roles* (e.g. subject, object, main verb, etc.) or *grammatical types* (e.g. noun, verb, adjective, etc.) of words. In most of the following studies, we focused on subject-verb (SV) and verb-object (VO) scenarios, i.e. when we have two parties, either S and V or V and O corresponding respectively to subject and verb or verb and object.

We then give each party a choice of inputs, corresponding to a choice of *word* to fill in their corresponding part-of-speech. In analogy to the Bell scenario described in Section 1.2, these choices of inputs can also be seen as local measurements. Let us look at an example and consider the lexicon:

$$\mathcal{L} = \{tap, pitcher, box, cabinet\} \quad (3.1)$$

and two parties S and V such that S can choose a subject for a verb chosen by V . Then, V is allowed to choose between the words of \mathcal{L} which are verbs, in this case, *tap* or *box*. Similarly, in the general case, S can choose between the set of words in \mathcal{L} which can be nouns, i.e. the whole of \mathcal{L} . However, for the empirical models described in the following sections, we will decide to manually restrict each party's input choices, e.g., by letting S choose between *pitcher* or *cabinet*.

Finally, each measurement outcome will correspond to possible interpretations of the input words. For instance, the word *pitcher* can have two possible interpretations:

- a. A large jug
- b. The position in batting sports (mostly baseball) in which the player delivers the ball to the batter

If all of the words in the lexicon are lexically ambiguous, the set of outcomes



(a) A jug



(b) A baseball pitcher

Figure 3.1: The two interpretations of the word *pitcher*

for each measurement is not the singleton, and each interpretation will come with probabilities. Furthermore, these probability distributions will be *dependent on their context*, meaning that the probability distributions are defined when all parties have made their choices of inputs (i.e. words). These probability distributions combined in different ways form empirical models that we will study in the next sections.

Quantum mechanics	Linguistics
Parties	Grammatical roles/types
Inputs/Measurements	Words
Outputs/Outcomes	Interpretations

Table 3.1: Analogy between quantum and linguistics scenarios.

To estimate those probability distributions, we created two datasets that we will refer to as the *corpus dataset* and the *human judgment dataset*. For both datasets, we started with a list of homonymous and polysemous nouns from [174, 156], and a list of homonymous and polysemous verbs from [146, 167]. The list of these ambiguous words can be found in Appendix D.1. To simplify calculations, we also restricted the choice of meanings (resp. senses) to two distinct meanings (resp. senses) per word. For example, even though the verb *tap* has multiple interpretations as a verb (e.g. touching something, secretly recording conversations, tap dancing, using up a resource, etc.), we decided to restrict to the following two meanings:

- a. Hit something gently, e.g. *tapping someone on the shoulder*
- b. Secretly listen or record what someone is saying using a device, e.g. *tapping phones*

3.1.1 The corpus dataset

Our first approach was to approximate these probability distributions using frequencies obtained from large corpora. To do so, we made use of two corpora, namely the British National Corpus (BNC) [43] containing 100 million words from a variety of sources and the `ukWaC` corpus [62] which contains more than 2 billion words obtained by crawling `.uk` web domains. Both corpora are part-of-speech annotated, but the semantic annotations had to be done by hand. Obtaining probabilities was then done as follows:

1. As we are interested in SV and VO phrases, we recorded every occurrence in the corpus where one of the ambiguous nouns (from our list in Appendix D.1) was the subject or object of one of the ambiguous verb (also in the list in Appendix D.1).
2. For each of these occurrences, we annotated the intended interpretation x_v, x_n of the verb $v \in \mathcal{L}$ and the noun $n \in \mathcal{L}$. For instance, if we found the SV phrase *the pitcher tapped* (i.e. $n = \text{pitcher}$ and $v = \text{tap}$) in the full sentence *The pitcher tapped his glove and glanced over at the runner on first base*, then we would have annotated it as:

$$\begin{aligned} x_{\text{pitcher}} &= \text{baseball player} \\ x_{\text{tap}} &= \text{hit something gently} \end{aligned}$$

3. For SV phrases, we then estimated the probability of the joint occurrence of the interpretations x_v, x_n in the context “ n is the subject of v ” as:

$$P[x_v, x_n \mid n \text{ subject of } v] = \frac{N((n, v) \mapsto (x_n, x_v) \wedge n \text{ subject of } v)}{N(n \text{ subject of } v)} \quad (3.2)$$

where N records the number of occurrences of each of the events, and $(n, v) \mapsto (x_n, x_v)$ correspond to the event where the noun n and verb v are interpreted as x_n and x_v respectively. Similarly, in VO phrases, we calculated the probability distributions of measuring the joint occurrence of the interpretations x_v and x_n

in the context “ n is the object of v ” as:

$$P[x_v, x_n \mid n \text{ object of } v] = \frac{N((n, v) \mapsto (x_n, x_v) \wedge n \text{ object of } v)}{N(n \text{ object of } v)} \quad (3.3)$$

The obtained dataset is available at [191].

Limitations of the dataset

This approach for collecting probabilities is intuitive, and the obtained probabilities are easily interpretable. In addition, large corpora are widely used in NLP and are easily (and freely) accessible. However, this approach also comes with some non-negligible drawbacks.

The most important one is the number of *joint occurrences* of two ambiguous words in a sentence, regardless of the grammatical relations imposed. These numbers were scarce, implying that the frequency obtained was not approaching the large number approximation of actual probabilities.

In addition, due to this small number of occurrences, not all possible combinations appeared in the corpora, and if they occurred, implausible interpretations in practice never appeared. For instance, we could easily imagine circumstances under which the phone conversations of a baseball pitcher would be recorded (e.g. if they were involved in a police inquiry). However, this meaning combination did not occur in either corpus.

An explanation for this phenomenon is that written texts and conversations are meant to be understood as efficiently as possible. Combining ambiguous words in a single sentence may increase the sentence’s overall ambiguity, making it unreadable.

A probabilistic argument could also explain this. Namely, the probability of occurrence of a di-gram is smaller than the probability of occurrence of either word in the di-gram. Given that some of the words on the list in Appendix D.1 did not occur very often to start with, e.g. the noun *pitcher* and the verb *to pen* only occurred 108 and 215 times respectively in the BNC, it would be unreasonable to expect the number SV or VO containing them to be high. Indeed, the phrase *the pitcher pens*, although meaningful, never occurred.

3.1.2 The human judgment dataset

To bypass the corpus dataset issues, we decided to ask *humans* to rate the plausibility of the ambiguous phrases. In particular, this allowed us to *choose* which phrases we wanted to study, and even highly unlikely meanings of phrases would be able to obtain a non-zero probability. In addition, fewer data points are necessary to get a reasonable estimate of the “real” probability distribution. Indeed, the judgment ratings of a single person are already approximations of the probability, whereas frequencies from corpora are dependent on the law of large numbers.

The data collection proceeded as follows:

1. We started with the same list of ambiguous nouns and verbs as per the corpus dataset (see Appendix D.1) and manually selected:
 - 50 (noun, verb) pairs consisting of both homonymous nouns and verbs;
 - 50 (noun, verb) pairs consisting of a homonymous noun and a polysemous verb;
 - 50 (noun, verb) pairs consisting of a polysemous noun is polysemous and a homonymous verb;
 - 50 (noun, verb) pairs consisting of both polysemous nouns and verbs.

The pairs selected were also checked to give a well-defined probability distribution for both SV and VO phrases, i.e. at least one of the meaning combinations will come with non-zero probability. These pairs can be found in the Appendix D.3.

2. We then split the 400 phrases into batches of 8 randomly chosen phrases containing only SV or VO phrases. We submitted the batches on the Amazon Mechanical Turk platform, where they were sent to workers to annotate.
3. 25 independent workers annotated each batch, and each worker was only allowed to annotate either an SV or a VO batch.

We then presented the workers with the following task:

4. A phrase (e.g. *the pitcher taps*) was shown to the annotator, who rated the plausibility of each of the meaning combinations as either:

- *Impossible* (score: 0)
- *Extremely unlikely* (score: 1)
- *Very unlikely* (score: 2)
- *Somewhat unlikely* (score: 3)
- *Neutral* (score: 4)
- *Somewhat likely* (score: 5)
- *Very likely* (score: 6)
- *Extremely likely* (score: 7)

5. For each phrase, we obtain a probability distribution from an annotator by normalising their score as follows:

$$P[(n, v) \mapsto (x_n, x_v)] = \frac{S(x_n, x_v)}{\sum_{\tilde{x}_n, \tilde{x}_v} S(\tilde{x}_n, \tilde{x}_v)} \quad (3.4)$$

Here, $S : \mathcal{I}_n \times \mathcal{I}_v \rightarrow \{0, \dots, 7\}$ is the function associating a score of an interpretation $(x_n, x_v) \in \mathcal{I}_n \times \mathcal{I}_v$. For instance, the set of scores corresponding to the SV phrase *the pitcher taps*:

Interpretation	<i>pitcher</i> \mapsto <i>jug</i> <i>tap</i> \mapsto <i>hit</i>	<i>pitcher</i> \mapsto <i>jug</i> <i>tap</i> \mapsto <i>record</i>	<i>pitcher</i> \mapsto <i>player</i> <i>tap</i> \mapsto <i>hit</i>	<i>pitcher</i> \mapsto <i>player</i> <i>tap</i> \mapsto <i>record</i>
Score	5	1	7	3

would have led to the probability distribution:

Interpretation	<i>pitcher</i> \mapsto <i>jug</i> <i>tap</i> \mapsto <i>hit</i>	<i>pitcher</i> \mapsto <i>jug</i> <i>tap</i> \mapsto <i>record</i>	<i>pitcher</i> \mapsto <i>player</i> <i>tap</i> \mapsto <i>hit</i>	<i>pitcher</i> \mapsto <i>player</i> <i>tap</i> \mapsto <i>record</i>
Probability	$\frac{5}{16}$	$\frac{1}{16}$	$\frac{7}{16}$	$\frac{3}{16}$

6. The probability distributions of all of the workers who annotated the same phrase were then averaged^[1].

An example of the task, as presented to the annotators, is illustrated in Fig. 3.2.

^[1]Note that that is the same as averaging the score and then normalising them

[View instructions](#)

... *the pitcher taps* ...

Question 1

How likely is the noun and verb in the above phrase interpreted as:

<p style="text-align: center;">Pitcher</p> <p style="text-align: center;">A large jug, e.g. <i>a ceramic pitcher</i></p>	<p style="text-align: center;">Tap</p> <p style="text-align: center;">Hit something gently (literal or figurative), e.g. <i>tap someone on the shoulder</i></p>
---	--

Impossible
 Extremely unlikely
 Very unlikely
 Somewhat unlikely
 Neutral
 Somewhat likely
 Very likely
 Extremely likely

Question 2

How likely is the noun and verb in the above phrase interpreted as:

<p style="text-align: center;">Pitcher</p> <p style="text-align: center;">A large jug, e.g. <i>a ceramic pitcher</i></p>	<p style="text-align: center;">Tap</p> <p style="text-align: center;">Secretly listen or record what someone is saying using a device (literal or figurative), e.g. <i>tapping phones</i></p>
---	--

Impossible
 Extremely unlikely
 Very unlikely
 Somewhat unlikely
 Neutral
 Somewhat likely
 Very likely
 Extremely likely

Question 3

How likely is the noun and verb in the above phrase interpreted as:

<p style="text-align: center;">Pitcher</p> <p style="text-align: center;">Baseball player that throws the ball to the batter, e.g. <i>the pitcher threw a lot of curve balls</i></p>	<p style="text-align: center;">Tap</p> <p style="text-align: center;">Hit something gently (literal or figurative), e.g. <i>tap someone on the shoulder</i></p>
---	--

Impossible
 Extremely unlikely
 Very unlikely
 Somewhat unlikely
 Neutral
 Somewhat likely
 Very likely
 Extremely likely

Question 4

How likely is the noun and verb in the above phrase interpreted as:

<p style="text-align: center;">Pitcher</p> <p style="text-align: center;">Baseball player that throws the ball to the batter, e.g. <i>the pitcher threw a lot of curve balls</i></p>	<p style="text-align: center;">Tap</p> <p style="text-align: center;">Secretly listen or record what someone is saying using a device (literal or figurative), e.g. <i>tapping phones</i></p>
---	--

Impossible
 Extremely unlikely
 Very unlikely
 Somewhat unlikely
 Neutral
 Somewhat likely
 Very likely
 Extremely likely

Figure 3.2: Example of a task seen by the Amazon Mechanical Turk workers.

3.2 On the quantum-like contextuality of ambiguous phrases

It is often said that natural language is “contextual”, notably in the context of lexical ambiguity. What is meant by that is that even though a single word, such as *pitcher*, can have multiple interpretations (see Section 3.1), it may have a single accurate interpretation given a context. For example, consider the following sentences (taken from [69]):

(1a) Being so elegantly designed, the *pitcher* pleased Mary.

(1b) Throwing so many curve balls, the *pitcher* pleased Mary.

In the context (1a), the only appropriate meaning of *pitcher* is *large jug*, whereas in (1b), the only appropriate meaning is *baseball player*.

However, the meaning of contextuality in quantum mechanics is different. Although, as in the linguistic sense, statistics of a system depend on their measurement contexts, the notion of contextuality is more complex than that. In quantum terminology, a system is said to be *contextual* iff the dependence of the statistics on the contexts is “essential” in the sense that it cannot be attributed to other factors (see Section 1.2). For instance, in the special case of non-locality, contextuality is observed if the statistics for the global measurement contexts cannot be explained entirely by its local behaviour. We are therefore interested in seeing whether the dependence of interpretation on (linguistic) context is also “essential” and whether we can witness quantum-like correlations between word interpretations.

Historically, contextuality in quantum mechanics has been proven via inequalities, usually referred to as *Bell inequalities*. These inequalities, however, depend on a crucial assumption of the system, namely that it is *no-signalling*.

We recall from Section 1.2, that a system is no-signalling iff the probability distributions agree on the intersections of their contexts, i.e. if all local sections are compatible. This condition is well motivated in the case of Bell-type experiments in Quantum Mechanics as the no-signalling property states that *information* cannot be transmitted faster than light (i.e. non-locally). However, obtaining perfectly no-signalling empirical models is, in practice, unfeasible due to the finite nature of experimental results and the imperfections of the experimental apparatus.

In addition, there is no fundamental reason why no-signalling should even apply in the case of ambiguities in natural language. In fact, the psycholinguistics literature would suggest that probability distributions arising from lexically ambiguous phrases *should be signalling*. For example, the probabilities associated with the different meanings of *pitcher* should be different in the phrase *ceramic pitcher* to the ones in the phrase *baseball pitcher*.

Several extensions of the notion of quantum contextuality have been proposed to account for signalling systems, including the Contextuality-by-Default (CbD) framework [52] (see Section 1.2.3), and the corrected Bell-inequalities approach [183] based on the sheaf-theoretic framework of contextuality [6] (see Section 1.2.2). In this section, we propose to apply these generalised frameworks to investigate the contextuality of lexically ambiguous phrases.

3.2.1 Cyclic models of rank 2

We start by discussing the simplest possible models, which only contain two words, a noun n and a verb v , and two different ways to combine them, in our case, as a subject-verb or verb-object phrase. In our previous analogy (see Table 3.1), this means that we have two parties, corresponding to grammatical types noun (N) and verb (V), each of them having a unique choice of input, but for which we can obtain two different probability distributions (one corresponding to SV and one corresponding to VO). These models are called *cyclic systems of rank 2* in the CbD literature [52, 111, 51].

Example 3.1. Consider the case where $n = \textit{pitcher}$ and $v = \textit{tap}$, taking the interpretations from Section 3.1. This choice of words leads to a valid SV phrase, namely *the pitcher taps ...*, and a VO phrase, namely *... taps the pitcher*. We can then associate the probability distributions with these two phrases^[2]:

(N, V)	(a., a.)	(a., b.)	(b., a.)	(b., b.)
$(\textit{pitcher}, \textit{tap})_{SV}$	5/16	1/16	7/16	3/16
$(\textit{pitcher}, \textit{tap})_{VO}$	17/22	0	15/22	0

This pair of probability distributions will be our empirical model for these contexts.

^[2]This is not an empirical model obtained from the corpus or human judgment dataset. The probability distribution for the SV phrase is taken from an example in Section 3.1, and the probability distribution for the VO phrase is taken from the corpus dataset.

Remark 3.2. This situation is similar to the question-order effect investigated by Wang et al. [196]. The work of [196] consisted of a series of behavioural experiments where participants were asked the same set of questions in different orders, and it was shown that the answers appeared to depend on the order of the questions that were asked. In [196], the authors argued that such experiments exhibit quantum-like contextuality. However, the authors of [49] demonstrated that the apparent contextuality was due to signalling.

We note here that these models are trivially non-contextual within the generalisation of the sheaf-theoretic framework of contextuality. This can be seen as follows. Let us start with a decomposition of a given (signalling) empirical model e :

$$e = \lambda \cdot e_{NS} + (1 - \lambda)e' \quad (3.5)$$

In the case of the models described above, in any no-signalling model e_{NS} satisfying (3.5), the probability distributions of both the SV phrase and the VO phrase collapse to a single probability distribution. Therefore, any of the no-signalling empirical models e_{NS} are trivially *non-contextual* (i.e. a global probability distribution exists and corresponds to the unique probability distribution in the empirical model).

However, these models can exhibit contextuality within the CbD framework. We will therefore focus on the CbD analysis of such cyclic systems of rank 2.

As first demonstrated in [194], it is indeed possible to find instances of linguistic empirical models of this form that exhibit CbD-contextuality. First, we found two contextual examples in the corpus dataset: empirical models where $n = \textit{boxer}$ and $v = \textit{adopt}$, and $n = \textit{pitcher}$ and $v = \textit{throw}$. These empirical models are depicted in Fig. 3.3. The degree of non-contextuality of both of these models can also be calculated to be respectively:

$$\text{NCNT2}(e_{(\textit{boxer}, \textit{adopt})}) = -\frac{1}{30} \quad (3.6)$$

$$\text{NCNT2}(e_{(\textit{pitcher}, \textit{throw})}) = -\frac{7}{30} \quad (3.7)$$

We recall from Section 1.2.3 that the fact that these numbers are negative shows that these models are contextual.

(N, V)	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$(boxer, adopt)_{SV}$	1/4	0	0	3/4
$(boxer, adopt)_{VO}$	0	29/30	1/30	0

(N, V)	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$(pitcher, throw)_{SV}$	0	2/3	1/3	0
$(pitcher, throw)_{VO}$	2/5	0	1/10	1/2

Figure 3.3: Empirical models of cyclic systems of rank 2 which were found to be CbD-contextual within the corpus dataset.

Remark 3.3. The study described in [194] provided the first instances of quantum-like contextuality in linguistic scenarios, which takes signalling into account. Previous work existing in the literature [32] have claimed to have violated Bell inequalities in natural language data. However, these did not assume the no-signalling condition and were later found to be non-contextual within the CbD framework [49].

Other contextual cyclic systems of rank 2 emerged from the human judgment dataset. These are found in Fig. 3.4b with their degree of contextuality.

(n, v)	$-NCTN_2$
(pitcher, throw)	0.233
(boxer, adopt)	0.033

(a) Corpus dataset

(n, v)	$-NCTN_2$
(file, admit)	0.232
(cabinet, reflect)	0.199
(volume, conduct)	0.111
(perch, file)	0.073
(plant, trap)	0.052
(press, file)	0.042
(swallow, admit)	0.021
(press, conduct)	0.011
(port, bill)	0.008
(organ, bill)	0.001

(b) Human judgment dataset

Figure 3.4: All of the found examples of CbD-contextual cyclic systems of rank 2 (sorted by degree of contextuality).

The presence of contextuality in quantum systems has some important consequences. From a foundational point of view, contextuality distinguishes between classical and quantum behaviours [22, 107]. From a computational point of view, it

is a resource that allows quantum advantages to arise [94, 3, 48]. The reader should refer back to Section 1.2.1 for a more detailed discussion.

In the case of the CbD-contextuality, this interpretation could be clearer. The witnesses of CbD-contextual in linguistics imply that the influence of the context over meaning selection is highly non-trivial and consists of a “truly contextual influence”. This finding concurs with the intuition that the context is the main factor in the interpretation of lexically ambiguous words.

Degrees of contextuality

In addition to finding out whether a given empirical model is contextual, it is also possible to calculate *how contextual* (or equivalently how non-contextual) a given empirical model is within the CbD framework^[3]. In this work, we will make use of NCNT2 defined in Section 1.2.3 to see how the levels of ambiguity of words can influence the *degree of contextuality*.

Results We calculated the degree of non-contextuality for the empirical models and classified them in terms of the ambiguity of their nouns and verbs (see Fig. 3.5).

We found that the contextuality mostly depends on the ambiguity of the verb. Specifically, the degree of non-contextuality is higher if the verb is polysemous. Equivalently, a rank 2 system will be more contextual whenever the verb is homonymous. Although we observed this trend in both the corpus and human judgment datasets, it was only statistically significant^[4] in the latter. The *t*-test comparing the data concerning homonymous and polysemous had *p*-value $p = 0.006$ in the human judgment dataset, as opposed to $p = 0.314$ in the corpus dataset.

In addition, we found no dependence on the levels of ambiguity of the nouns and the degree of (non-)contextuality in either dataset. The *p*-values were $p = 0.38$ and $p = 0.15$ for the corpus and the human judgment dataset, respectively.

Discussion The conclusions one can draw from the degree of (non-)contextuality are hindered by the need for a more operational interpretation of CbD-contextuality.

^[3]It is also possible to calculate some measures of contextuality in the sheaf-theoretic framework, the contextual fraction CF being the most prominently used [3, 4]. We are, however, not making use of them in this thesis.

^[4]The definition of statistical significance is standardly defined as follows. A feature is said to be statistically significant iff the associated *p*-value is less than 0.05.

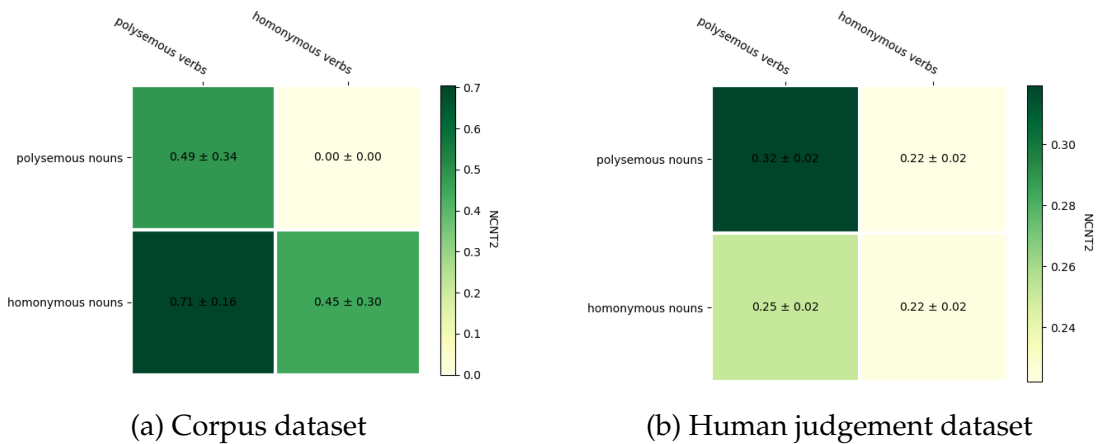


Figure 3.5: Mean NCNT2 depending on whether the noun and verb were polysemous or homonymous. The errors quoted are the standard errors of the means.

Therefore, the interpretation of the result is not completely clear.

A possible interpretation of finding contextual witnesses is the following. Recall that, in the case of cyclic systems of rank 2, the choices of contexts are whether the pair of words is found in an SV or VO phrase. This result suggests that a verb would use its context more whenever its possible interpretations are unrelated.

This conclusion is consistent with the fact that homonymous verbs use the knowledge of both their subject and object in their first disambiguation stage. In contrast, polysemous verbs will use information from a broader context (if available) to be disambiguated.

3.2.2 Cyclic models of rank 4

We have also studied a larger type of empirical model: cyclic systems of rank 4. These models are analogues of the (2,2,2)-Bell scenarios, i.e. scenarios that consist of 2 parties, each of these parties having 2 choices of inputs, and each local measurement having at 2 possible outcomes. In these models, we take the parties to represent specific dependency relations, e.g. main verb, subject, or object. In particular, we will focus on SV models (with parties S and V) and VO models (with parties V and O). In each model, the parties will choose between two words of the correct grammatical type (i.e. nouns for parties S and O and verbs for V parties). Each word will then have two different possible interpretations.

Example 3.4. Let's consider an example of a VO model where V is allowed to choose

between the verbs in $\{tap, box\}$ and O is allowed to choose the object of these verbs in the set $\{pitcher, cabinet\}$. The interpretations of *tap* and *pitcher* are taken to be the ones from Section 3.1. In addition, the interpretations of *box* and *cabinet* are respectively:

- a. To put something in a box, e.g. *boxing up clothes and books*
- b. To practice the sport of boxing, e.g. *He boxed professionally for years*

and:

- a. A small group of the most important people in government, e.g. *a cabinet minister*
- b. A piece of furniture with shelves, cupboards, or drawers, e.g. *a glass-fronted cabinet*

Then, from the corpus dataset, we obtained the following empirical model:

(V, O)	(a., a.)	(a., b.)	(b., a.)	(b., b.)
(tap, pitcher)	17/22	15/22	0	0
(tap, cabinet)	1/21	3/7	11/21	0
(box, pitcher)	3/4	1/4	0	0
(box, cabinet)	3/7	10/21	2/21	0

These types of models have the potential to exhibit contextuality in both the Contextuality-by-Default framework and the extension of the sheaf-theoretic framework for corrected Bell inequalities. However, none of the empirical models obtained in the corpus or human judgment dataset were contextual (in either framework).

Regarding the CbD framework, this could be explained as the probability of obtaining a contextual model decreases as the systems get bigger. We can estimate these probabilities by random sampling from a uniform distribution, and the likelihood of obtaining a CbD-contextual cyclic system of rank 2 is about 17% and drops to about 0.01% for cyclic systems of rank 4. In addition, violations of the corrected Bell inequalities of [183] is a stronger condition than the CbD notion of contextuality (see Section 1.2.3). Therefore, obtaining a contextual model within this framework is also quite unlikely.

Even though we haven't found any contextual cyclic system of rank 4, this does not mean that *no such system* is contextual. A larger scale experiment will be needed to obtain witnesses of contextuality in these types of models – we leave this to future work.

Degrees of contextuality

Although we have not been able to find contextual witnesses in cyclic systems of rank 4, we can study the degrees of (non-)contextuality of the obtained empirical models, as we did for cyclic systems of rank 2.

Methods We want to know how the levels of ambiguity of words of different syntactic roles (i.e. subject, verb, or object) influence the degree of contextuality of the respective empirical models. In addition, each empirical model can have:

- 0 polysemous verbs & 2 homonymous verbs;
- 1 polysemous verb & 1 homonymous verb;
- 2 polysemous verb & 0 homonymous verbs;

and similarly for subjects and objects. Therefore, we will classify the SV and VO models in terms of their numbers of homonymous verbs, subjects, and objects^[5]. We will mostly focus this analysis on the human judgment dataset since the corpus dataset did not have empirical in all categories. For example, the instances recorded did not lead to any SV cyclic system of rank 4 with two polysemous subjects and two homonymous verbs.

In addition, we are interested in the monotonic relations between the number of homonymous words (of each type) and the degree of contextuality, i.e. whether contextuality increases or decreases as the number of homonymous verbs, subjects, or objects increases. On the other hand, the existence of non-monotonic relations between these quantities does not lead to easily interpretable results. For example, it is unclear what it would mean for the contextuality to be higher whenever we can choose between a polysemous and a homonymous verb. Hence, we will make use

^[5]We could have equivalently chosen to classify them in terms of their number of polysemous verbs, subjects, and objects. However, the adopted convention fits our intuition that homonymous words are, to some extent, “more ambiguous” than polysemous ones.

of Spearman's rank correlation coefficient ρ , which will assess whether a monotonic relation exists between two random variables, one being the number of homonymous verbs, subjects, or objects and the other being the degree of non-contextuality NCNT2.

Results The values of the degrees of non-contextuality for each class of empirical models can be found in Fig. 3.6. An ANalysis Of Variance (ANOVA) first shows that the degrees of non-contextuality are statistically different across the different categories ($p < 10^{-4}$ for SV models and $p = 0.005$ for VO models).

In addition, we observe that in SV models, the degree of non-contextuality increases as the number of polysemous verbs and subjects increases. This can be verified using Spearman's correlation coefficient ρ , which was $\rho = -0.27$ with associated p -value $p < 10^{-6}$ for the correlation with respect to the number of homonymous verbs, and $\rho = -0.20$ with $p = 0.0002$ for the correlation with respect to the number of homonymous subjects. In both cases, the negativity of the ρ 's shows that NCNT2 decreases as the number of homonymous verbs and subjects increases. In addition, the fact that p -values are both < 0.05 shows that we are more than 95% confident that these coefficients are different from 0 (i.e. a correlation exists with a 95% confidence).

In VO models, these trends are much milder and, in fact, not statistically significant. We can see that the degree of direct influence decreases as the number of homonymous verbs decreases ($\rho = -0.11$, $p = 0.053$). However, no monotonic correlation is found with respect to the number of homonymous objects ($\rho = 0.03$, $p = 0.52$).

Discussion As in cyclic systems of rank 2, the main factor influencing the (non-)contextuality of the systems is the levels of ambiguity of the verbs. Indeed, we have already shown that the degree of contextuality increases as the number of homonymous verbs increases. This reinforces the intuition that homonymous verbs use their arguments (here, the context) more intrinsically than their polysemous analogues.

In addition, the same applies to homonymous nouns in SV phrases. Namely, homonymous nouns would lead to a higher amount of "true" contextuality in the obtained probability distributions.

The fact that this occurs in SV models only (and not in VO models) suggests that this finding relates to the position of the disambiguating context of homonymous

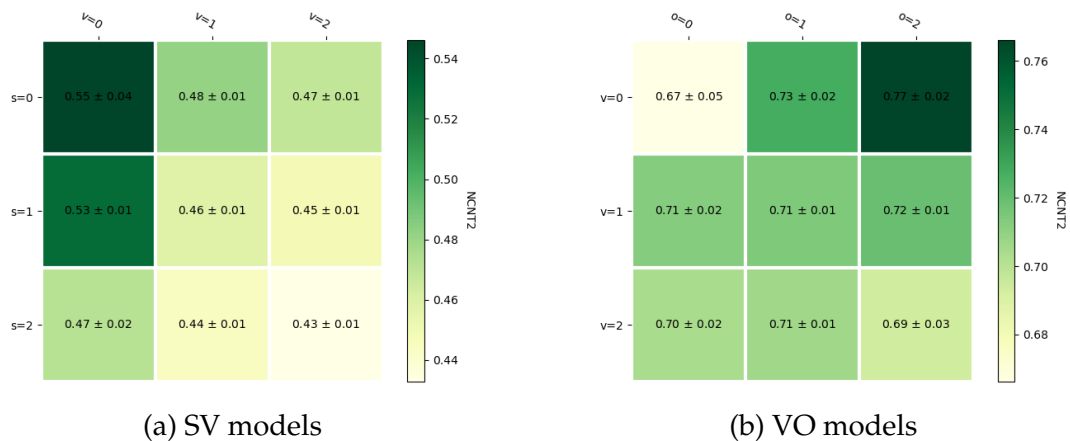


Figure 3.6: Averages of the NCNT2 as a function of the number of homonymous verbs, subjects, and objects.

nouns. In particular, a slowdown in the reading time of homonymous nouns has been observed when the disambiguating context is found after the target noun [69], but this slowdown was lesser for polysemous nouns. Hence, in the case of SV phrases, where the disambiguating context can only be the verb (situated after the noun), the observed high degree of contextuality suggests that homonymous nouns use their context in a less trivial way when the disambiguating context is found after it.

3.3 Degrees of signalling and the levels of ambiguity

In the previous section, we saw that signalling is the main obstacle to studying contextuality in natural language data. However, the presence of signalling is not in itself a weakness of natural language data. In fact, in most linguistic studies of lexical ambiguity, the fact that the interpretation of ambiguous words changes with the context, i.e. signalling, is the focus point. Here, we propose to study the amount of signalling present in the different empirical models and see what conclusions can be drawn.

Two ways of quantifying signalling are available to us, namely SF coming from the sheaf-theoretic framework of contextuality, and Δ from the CbD framework. In addition, the latter can be split with respect to the individual input choices. There-

fore, we can study the amount of signalling coming from a specific choice of word or, equivalently, how different the probability distributions of a given word are in the different contexts it is found in. We will then study the correlations between these different quantities and the levels of ambiguity of words in cyclic systems of rank 2 and 4.

Remark 3.5. Although the sheaf-theoretic and Context-by-Default frameworks are essential in the definition of the signalling fraction and the degrees of direct influence respectively, the mathematical machinery they employ are only used implicitly in this Section.

Remark 3.6. The results of this section may appear challenging to interpret and reason about, most of all because some results were verified in one dataset but not the other. In addition, it is not common in the linguistic literature to study phrases where more than one word is clearly (lexically) ambiguous. In particular, not much is known about what happens if the context of an ambiguous word is itself ambiguous.

3.3.1 Cyclic systems of rank 2

We start with the cyclic systems of rank 2, which we recall contains a noun/verb pair and two different contexts, SV and VO.

Overall degrees of signalling

We start by looking at the total degree of direct influence Δ , as well as the signalling fraction SF, both of which measure how signalling the whole system is (we also recall from Section 1.2.3 that these two quantities are not unrelated). Then, as we did in the analysis of the degree of contextuality in the previous section, we partition both our datasets in terms of the levels of ambiguity of the nouns and the verbs.

Results The overall degrees of signalling for all of the different classes of empirical models are shown in Fig. 3.7. We observe the same trend in both of these datasets, namely that the overall signalling of the system increases as the number of polysemous words increases. An ANOVA reveals a statistical difference between SF or Δ and the different classes of models in *the human judgment dataset only*. The p -values

were $p = 0.015$ for Δ and $p < 10^{-9}$ for SF in the human judgment dataset, as opposed to $p = 0.98$ and $p = 0.70$ for Δ and SF respectively in the corpus dataset.

In the human judgment dataset, we can verify using a t -test that Δ and SF are higher for nouns with multiple senses ($p = 0.04$ and $p = 0.02$ respectively). Similarly, Δ and SF are also higher for verbs with multiple senses ($p = 0.02$ and $p = 0.03$ resp.). No statistically significant trend could be found in the corpus dataset.

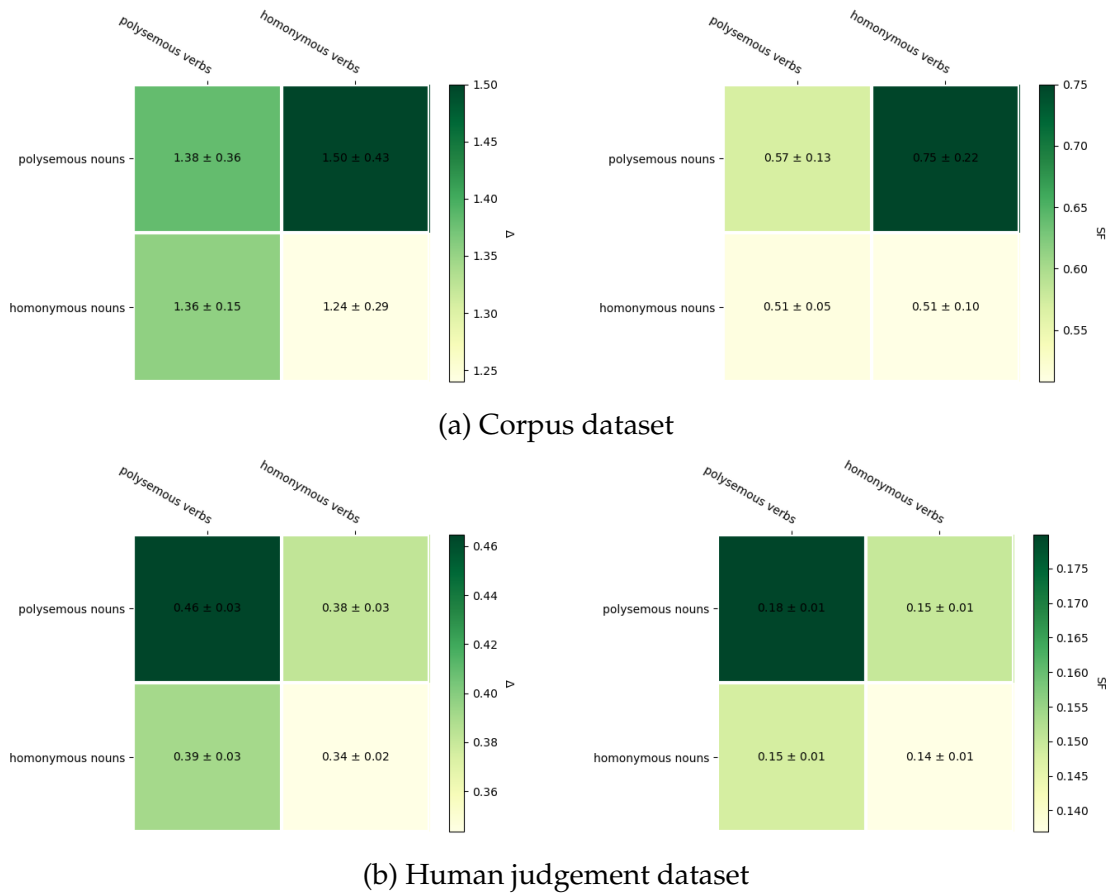


Figure 3.7: Averaged Δ and SF of the cyclic systems of rank 2, depending on whether the noun and verb were polysemous or homonymous. The errors quoted are the standard errors of the means.

Discussion This phenomenon could be explained using the theory of *underspecification*, in which the interpretation of polysemous words is essentially *created* from its context, whereas meanings of homonymous words are *selected* using contextual information [71] (see also Section 2.1.2). Hence, assuming that the SV and VO contexts are unrelated, this would imply that there is potential for having different in-

interpretations of the same polysemous word. On the other hand, one may argue that its context window is too small for homonymous words to swing widely from one meaning to the other.

Individual degrees of signalling

Given the above interpretation of the results, we also expect the individual degrees of direct influence to follow the same tendency.

Results This is, however, not quite verified at the level of individual degrees of direct influence. In particular, in the human judgment dataset, the degree of direct influence from the verb was higher whenever the verb was polysemous, which is consistent with our hypothesis. On the other hand, the degree of direct influence from the noun was higher whenever the noun was *homonymous*. In both cases, the observed effect was relatively small, and the difference in individual direct influence was 0.06 for verbs of different levels of ambiguity and 0.08 for nouns of different levels of ambiguity. In both cases, these differences, however small, were still found to be statistically significant (with respective p -values $p = 0.006$ and $p < 10^{-4}$).

In the corpus dataset, the reverse trend is observed (i.e. homonymous verbs and polysemous nouns had, on average, higher degrees of direct influence). However, none of the differences were statistically significant ($p = 0.27$ for verbs and $p = 0.78$ for nouns).

Discussion Due to the size of the effect and the fact that datasets did not agree on the findings, we could conclude that such a difference may be due to statistical fluctuations. However, if these effects were in fact “true”, this would imply that some more complex mechanism occurs in the disambiguation process of both nouns and verbs.

On the disambiguation windows of ambiguous verbs

The corpus dataset In [195], we showed that the *proportion* of the direct influence coming from the verb was statistically significantly higher for homonymous verbs than for polysemous verbs (see Fig. 3.8) in the corpus dataset. This fact was attributed to the difference in disambiguation windows in homonymous and polysemous verbs.

Indeed, we recall that the first disambiguation stage for homonymous verbs happens as soon as all of its arguments are known. In contrast, the reader only selects the senses of polysemous verbs at the end of the phrase or sentence.

Now, as in cyclic systems of rank 2, we are studying the difference in distributions between subject-verb and verb-object contexts. We can expect differences in the distributions of the interpretations of homonymous verbs, as we are precisely within this first disambiguation stage.

The fact that we did not observe any such effect for nouns was explained by the fact that the changes of contexts studied remain within the same disambiguation window for both polysemous and homonymous nouns.

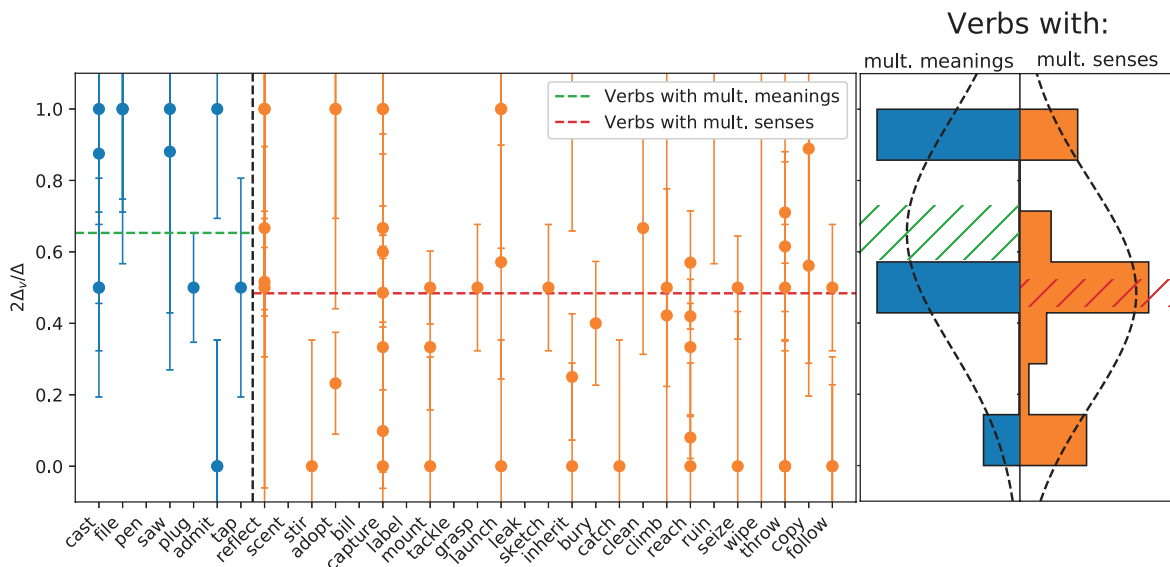


Figure 3.8: Relative contributions of the verb content to the overall direct influence given different levels of ambiguity for the verb or the noun. The left-hand figures correspond to the contributions of the verb; averages for each level of ambiguity are shown with dotted lines. The right-hand figures correspond to the distributions of these data points. The hatched area depicts the 66%-confidence intervals for the means. The fitted normal distributions are also plotted.

The human judgment dataset In the human judgment dataset, this relation was not verified. In fact, we observed the opposite, namely that the proportion of direct influence from polysemous verbs was greater than the one for homonymous verbs, but this was not statistically significant ($p = 0.22$).

On the other hand, the degree of direct influence from the verb was, on average,

higher than the degree of direct influence from the noun ($p = 0.003$ for a related t -test). This observation is consistent with the fact that verbs need their arguments to be disambiguated, whereas nouns do not. Hence, the variations in the interpretations are greater between the two contexts (SV or VO) compared to the variations in the interpretations of the nouns. However, we would have expected from the previous discussion that homonymous verbs would have higher degrees of direct influence, which is not the case in the human judgment dataset (see above).

In addition, we should also note that this is *not* at all observed in the corpus dataset ($p = 0.51$). This lessens the findings of [195], but, on the other hand, offers alternative evidence that readers do not disambiguate verbs and nouns in the same way and that the presence of the arguments of the verb is essential in their disambiguation process.

On the ambiguity of the context

Lastly, some cross-effect between the ambiguity of one word and the degree of direct influence of the other has been observed in both datasets.

Results In the human judgment dataset, Δ_v was higher whenever the verb was combined with a polysemous noun (with an average difference of 0.13 and $p < 10^{-9}$). We also observed a similar effect in the corpus dataset, but the effect size is much smaller (average difference of 0.014) and not statistically significant ($p = 0.95$).

In addition, Δ_n was slightly higher whenever the noun in the empirical model was combined with a polysemous verb (average difference of 0.004 in the human judgment dataset and 0.27 in the corpus dataset). However, in this case, none of the observed differences were statistically significant ($p = 0.82$ in the human judgment dataset and $p = 0.21$ in the corpus dataset, respectively).

Discussion The interpretation of this fact would be related to something that has yet to be studied in the psycholinguistic literature, namely, what happens if the context itself is ambiguous? This result suggests that if the context is polysemous or underspecified, the interpretation of a target word becomes more variable. In contrast, if the context can have several unrelated interpretations, then the interpretation of the target word is also more defined. This will be made more transparent

and intuitive in Section 3.4 when studying the *causality* of the systems instead of its signalling.

3.3.2 Cyclic systems of rank 4

We now look at the degrees of signalling of the cyclic systems of rank 4. As in Section 3.2.2, we do not have enough data in the corpus dataset to cover all possible combinations of the number of polysemous and homonymous subjects, verbs, and objects. Hence, we will shift our focus to the human judgment dataset only.

Remark 3.7. In terms of notation, we will denote as Δ_A , $A \in \{S, V, O\}$ for the total amount of direct influence coming from the two subjects, verbs, or objects (i.e. the different parties), and Δ_a for $a \in \{s, v, o\}$ for the individual direct influence coming from a particular choice of subject, verb or object (i.e. the individual choices of inputs). For instance, in the SV empirical model with inputs $\{s_1, s_2, v_1, v_2\}$, we would have:

$$\begin{aligned}\Delta_S &= \Delta_{s_1} + \Delta_{s_2} \\ \Delta_V &= \Delta_{v_1} + \Delta_{v_2}\end{aligned}$$

and:

$$\Delta = \Delta_S + \Delta_V = \sum_{w \in \{s_1, s_2, v_1, v_2\}} \Delta_w$$

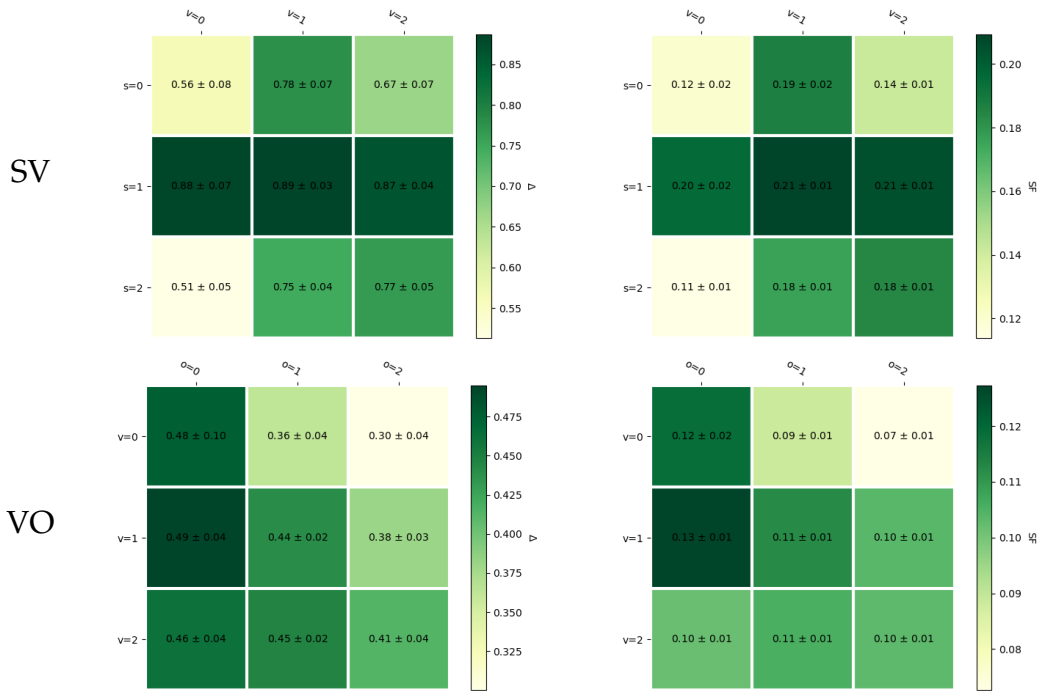
Subject-Verb vs. Verb-Object

We start by looking at the difference in signalling between SV and VO models.

Results The degree of signalling, measured by both the signalling fraction SF and CbD measure Δ are (statistically) significantly higher for SV models compared to the VO models (the average difference was 0.38 for Δ and 0.08 for SF, with respective p -values $p < 10^{-52}$ and $p < 10^{-40}$).

Discussion This suggests that the interpretations of VO phrases are easier to obtain, and their interpretations are more well-defined than those of SV phrases. This concurs with the usual grouping of VO compounds as VPs, whereas subject-verb

does not generally correspond to anything special, for example, in context-free grammars [76].



(a) SF and Δ as a function of the number of homonymous subjects, verbs, and objects. The errors quoted are the standard errors of the mean.

		R	p
Δ	Subject	0.0170	0.7607
	Verb	0.0329	0.5568
SF	Subject	0.0466	0.4045
	Verb	0.05353	0.3383

(b) SV

		R	p
Δ	Verb	0.13755	0.01349
	Object	-0.1560	0.005
SF	Verb	0.0650	0.2444
	Object	-0.1289	0.0207

(c) VO

Table 3.2: Analysis of the monotonicity of the amount of signalling and the number of homonymous words in SV and VO empirical models.

Verbs vs. Nouns

As in the cyclic system of rank 2, the degree of direct influence from the verb, Δ_V , was found to be higher than the direct influence from both the subject (Δ_S) and the object (Δ_O); in SV models, the average difference was 0.1961 with associated p -value $p < 10^{-20}$, and in VO, the average difference was 0.08 with associated p -value

$p < 10^{-13}$. As before, we justify this finding as the verbs are, in general, not fully disambiguated at these stages, and therefore, their interpretation is more variable.

Degrees of signalling for homonymous and polysemous words

Ambiguity of the objects We observe that in VO models, the degrees of signalling (both SF and Δ) increase as the number of polysemous objects increases. The obtained Spearman's correlation coefficients are $\rho = 0.16$ for Δ and $\rho = 0.13$ for SF, and the respective p -values are $p = 0.005$ and $p = 0.02$.

This finding is consistent through all of the fine-grained measures of direct influence. Indeed, Δ_O was increasing the more polysemous objects were in the empirical models ($\rho = 0.14$, $p = 0.01$), and the individual degrees of direct influence Δ_o were also higher for polysemous than homonymous objects (average difference of 0.02, and associated p -value $p = 0.03$).

This resonates with the theory of underspecification which happens in polysemous words. In other words, since noun senses are more dependent on their context than noun meanings, it makes sense that the interpretation of a polysemous word would vary more than the interpretation of a homonymous word.

Ambiguity of other dependencies The effect described above is only observed in objects in VO models. In the other cases, the different degrees of signalling did not appear to be (monotonically) related to the numbers of polysemous/homonymous subjects or verbs (see Table 3.2), or if they "exist", their size is very small.

We will attribute this lack of relations between the ambiguity of the different words and their degrees of direct influence to other deciding factors in the process of their disambiguation. For instance, the disambiguation of the verbs in both SV and VO models will be affected not only by their own levels of ambiguity but also by the ambiguity of their arguments (i.e., their subject or objects). Similarly, the subject could also be affected by their context, either by the position or ambiguity of the context.

On the ambiguity of the context

As for cyclic systems of rank 2, it is possible to find some relations between the degree of ambiguity of the context and the degree of direct influence of a given

target word, although these correlations were very moderate.

Results In both SV and VO models, the degree of direct influence of the verbs Δ_V increased as the number of polysemous arguments (subject or object) increased. Spearman's correlation was $\rho = 0.08$ in VO systems and $\rho = 0.005$ in SV systems, and neither were statistically significant ($p = 0.17$ and $p = 90.3$ respectively). On the other hand, Δ_S and Δ_O increased as the number of homonymous verbs increased ($\rho = 0.15$ and $p = 0.006$ in the case of Δ_S and $\rho = 0.17$ and $p = 0.002$ in the case of Δ_O).

Discussion At first sight, this appears contradictory to the findings in rank 2 systems, where we recall that Δ_n increased as the number of polysemous verbs increased. However, one crucial difference here is that context is more "fixed" in SV and VO models, whereas it is not in cyclic systems of rank 2 previously described. For instance, consider an SV model. Taking the subject to be the target word, the context $_v$ is fixed in the two contexts $s_1 v$ and $s_2 v$. In contrast, in cyclic systems of rank 2, the two contexts of a noun would be $_v$ and $v _$, which are fundamentally different.

Moreover, the influence of the ambiguity of the verb of their subject and object could be understood as follows. The disambiguation of homonymous verbs first starts when the arguments are established. Hence, the reader will begin disambiguating both the verb and the noun, which creates some interaction between the choice of meaning of the verb and the subject/object. This makes the interpretation of the subject/object more uncertain. On the other hand, if the verb is polysemous, the meaning of the nouns will be more well-defined, as the disambiguation of the verb itself will be delayed to the end of the sentence (and hence beyond the scope of our experiments).

3.3.3 Discussion of the results

The signalling property of empirical models provides insight into the mechanisms at the heart of the disambiguation process. However, our analysis is made difficult and not easily interpretable from how empirical models are formed. Indeed, to have non-trivial empirical models, all of the inputs must be ambiguous. Yet, this adds complexity to the study as the distinction between (linguistic) context and target

word is symmetric, i.e. a word can be both a target-word and a context-word. In addition, the notion of signalling is also bidirectional, i.e. we can only see whether a dependence exists between two variables A and B , not whether A influences B or the other way around. In the next section, we will remedy these problems by examining the causal influences between the different parts-of-speech.

3.4 The causality of the disambiguation process

In the previous section, we have seen that the signalling property of natural language systems is not necessarily a hindrance but can give us some insight into human behaviour when disambiguating lexically ambiguous words and phrases. In this section, we go one step further and study the *structure* of the signalling present in natural language data. To do so, we make use of the extensions of the sheaf-theoretic framework of contextuality to account for *causality* [79, 80, 5] (see Section 1.2.2). In this line of research, we are interested in the *direction* of signalling. For instance, does the disambiguation order follow the reading order, i.e. is disambiguation purely incremental, or is backtracking necessary?

Remark 3.8. In this perspective, we can see the event of choosing the input as reading a new word, associating an outcome to a choice of measurement will then correspond to the disambiguation step. Hence, having a causal order that does not follow the linear ordering of the words in a given sentence does *not* mean that the reader reads the words in a different order, but rather that the *understanding* process is not instant (i.e. does not follow the reading order).



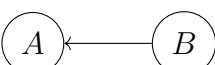
As for the no-signalling property, obtaining statistics that are fully compatible with a given causal order is not feasible in practice. Therefore, instead of calculating the no-signalling fraction $\text{NSF} = 1 - \text{SF}$ (i.e. the amount of the empirical model which is compatible with a perfectly no-signalling system), we can calculate the *causal fraction* introduced in Section 1.2.2 which measure the amount of the empirical model which is consistent with a given causal order [79]. In addition, similar to the previous sections, we will focus on the simplest non-trivial scenarios to minimise the calculations. Here, the smallest non-trivial system we can consider is similar to the

(2,2,2)-Bell scenarios (or cyclic systems of rank 4). Indeed, in the case where only 1 party is present, there is a unique compatible causal order: the party A can influence itself, and every such system would be trivially consistent with it. Similarly, if we have 2 parties, but at least one party has a unique choice of input, then this scenario reduces, without loss of generality, to the one-party case. Finally, if only 1 outcome is possible for all of the measurements, then all of the measurements are deterministic, and the analysis becomes once again trivial.

Remark 3.9. This work does not assume that parties are spacelike separated entities. Instead, the notion of party corresponds to entities isolated in time or space. For instance, we could consider a single person in a lab performing a sequence of 3 sequential measurements to count as 3 different parties.

The systems that we will consider are, as in Sections 3.3 and 3.2, SV and VO systems, where the two parties are either S and V or V and O and the interpretations of input and outcomes are left unchanged. In particular, we here focus on *definite causal orders*, i.e. causal orders represented by direct acyclic graphs [142]. In such graphs, the nodes correspond to random variables, which can have inputs and outputs. The directions of the arrows represent the causal relations, e.g. if $A \rightarrow B$, then this means that the input of A can influence the output of B , and the absence of arrows shows the independence of random variables. Finally, the acyclicity condition ensures that a given event cannot influence its past.

In a two-party system like ours, say with parties A and B , there are only three possible definite causal orders, namely:

1. 
2. 
3. 

The situation of 1 corresponds to our familiar no-signalling scenarios. The cases 2 and 3 respectively represent situations where the party A can influence B and where the party B can influence A . We will denote the causal fractions associated with causal orders 1, 2 and 3 as respectively $\text{CausF}_{NS} = \text{NSF}$, $\text{CausF}_{A \rightarrow B}$ and $\text{CausF}_{B \rightarrow A}$.

Remark 3.10. The causal orders $A \rightarrow B$ and $B \rightarrow A$ are not mutually exclusive. Therefore, we could have a model which is highly consistent with both causal orders separately.

Remark 3.11 (On the no-signalling causal order). A system with parties A, B is said to be no-signalling iff it is compatible with both causal models of the form $A \rightarrow B$ and $B \rightarrow A$. Therefore, it is stronger than causal orders $A \rightarrow B$ and $B \rightarrow A$. Therefore, in terms of the causal fractions, this means that:

$$\text{NSF} \leq \min(\text{CausF}_{A \rightarrow B}, \text{CausF}_{B \rightarrow A}) \quad (3.8)$$

We are investigating which of these causal orders is the most relevant, i.e., explains most of the system's statistics. We do so by comparing the different causal fractions obtained for the different causal order, notably 2- 3.

Calculating the causal fraction of an arbitrary model is not straightforward, as it requires solving an optimisation problem. However, given the specific form of the models we are considering, the causal fractions for each causal order can be calculated efficiently.

Proposition 3.12. *In a (2,2,2)-Bell-type scenario with parties A and B , the causal fraction is given by:*

$$\text{CausF}_{A \rightarrow B} = \min_{a \in \{a_1, a_2\}, o \in \{0,1\}} 1 - |e_{(a,b_1)}|_A(o) - e_{(a,b_2)}|_A(o)| \quad (3.9)$$

The causal fraction for the $B \rightarrow A$ causal order can be obtained by applying the formula to a relabelled empirical model.

This proposition's proof can be found in Appendix C.4.

Through initial calculations, we have found that the empirical models obtained in the corpus dataset were too sparse, and therefore did not lead to any conclusive result. In the rest of this work, we will focus on the causal analysis of the human judgment dataset, as their probability distributions are, on the whole, of better quality.

3.4.1 The direction of signalling in SV and VO models

We start by looking at the compatibility of our data with the different causal orders described above.

Results The causal fractions obtained for the SV and VO models are shown in Fig. 3.9a and 3.9b, respectively.

The data reveals that SV phrases are predominantly compatible with the $S \rightarrow V$ causal order. Indeed, all of the models have a causal fraction $\text{CausF}_{S \rightarrow V} > 0.7$ (the causal fractions $\text{CausF}_{S \rightarrow V}$ is on average 0.89), and both the $V \rightarrow S$ and the non-signalling fractions achieve lower causal fraction values where the causal fractions are on average 0.83 and 0.19 respectively, see also Fig. 3.9c.

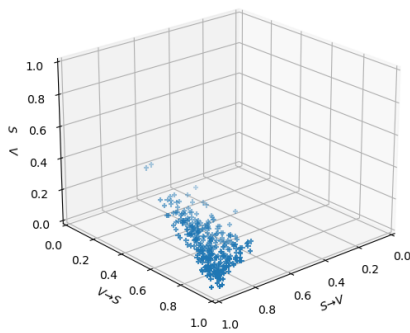
Similarly, the VO models achieve a causal fraction with the $O \rightarrow V$ order higher than 76% (and 0.93 on average), and the other causal fractions reach significantly lower scores ($\text{CausF}_{V \rightarrow O} \sim 0.91$ and $\text{SF} \sim 0.11$), see Fig. 3.9d. We note that in the case of the VO models, this suggests that the disambiguation order in these phrases is *opposite to the reading order* (assuming the standard active voice SVO structure of English).

Discussion These results show that the interpretation of ambiguous verbs is more affected by the choices of arguments (i.e. subject or object) rather than the other way around. This result is also validated by the research in psycholinguistics, which has indeed shown that the reader delays the disambiguation process until the arguments of the verb are known in the case of a homonymous verb and until the end of the phrase or sentence in the case of polysemous verbs [146].

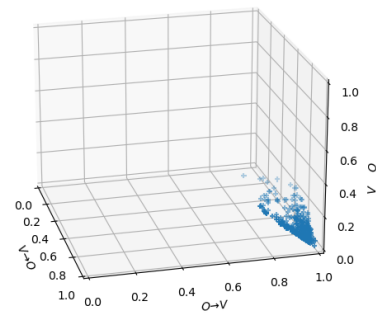
In addition, the causal fractions of VO models are generally higher and less variable (i.e. smaller standard deviation) than those obtained in SV models, see Fig. 3.10. This may suggest that the disambiguation process for VO ambiguous phrases is more straightforward than for SV ones.

3.4.2 Models with different levels of ambiguity

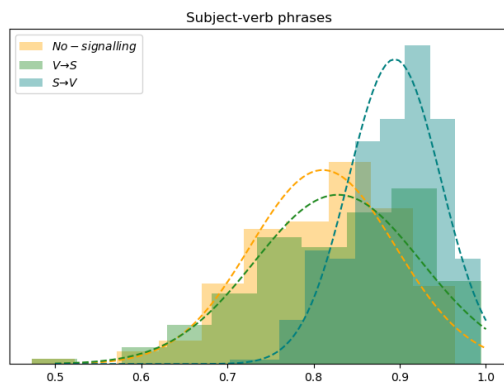
We now investigate whether we can observe a difference in behaviour between homonymous and polysemous words. We first recall that eye-tracking data suggests that readers tend to disambiguate polysemous words much later than their homonymous counterparts by selecting an underspecified interpretation instead of committing to a single sense. In SV models, this should result in having a higher causal fraction associated with $S \rightarrow V$ whenever the verbs are polysemous and the nouns are homonymous, as this case would delay the disambiguation of the verbs (compared to average), and make the disambiguation of the homonymous noun



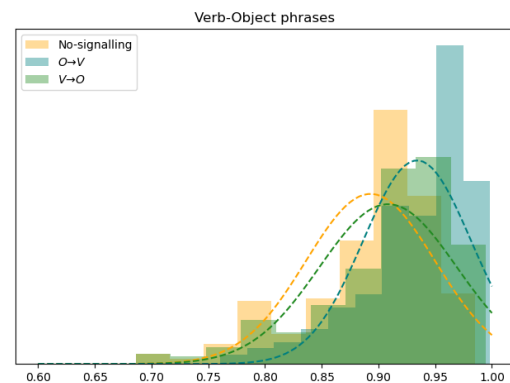
(a) Scatter plot of the different causal fractions for SV empirical models



(b) Scatter plot of the different causal fractions for VO empirical models



(c) Histogram of the recorded causal fractions for SV empirical models



(d) Histogram of the recorded causal fractions for VO empirical models

Figure 3.9: Distributions of the causal fractions

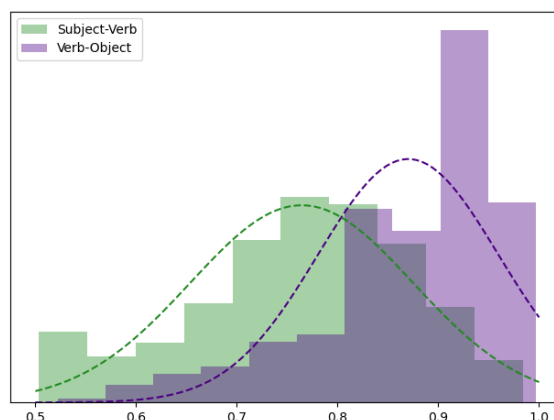


Figure 3.10: Comparison of the distributions of the $S \rightarrow V$ and $O \rightarrow V$ causal fractions.

faster than the average scenario. Similarly, in VO models, we expect the causal fraction associated with $O \rightarrow V$ to be larger whenever the verbs are polysemous and the nouns are homonymous. To verify this hypothesis, we study the correlation between the causal fractions and the number of homonymous/polysemous verbs and nouns in the different models.

Results & Discussion We started by looking at the effect of the ratio of homonymous/polysemous words in the empirical models to the causal fractions. From the above description, this ratio shouldn't have a significant effect. To check this, we calculated the Spearman ρ coefficients of the causal fractions and the number of homonymous words. We did not observe any apparent correlation in VO models ($\rho = 0.009$, $p > 87\%$). In SV models, we only observed a mild (but statistically significant) effect. In that case, we observed that the more polysemous the words in a model, the higher the $S \rightarrow V$ causal fraction ($\rho = 0.15$, $p < 0.7\%$). These results are depicted in Fig. 3.11 and are overall consistent with our hypothesis.

We then sort our empirical models in terms of their number of homonymous and polysemous *nouns*, and subsequently in terms of their number of homonymous and polysemous *verbs*. The observed distributions of the causal fractions are depicted in Fig. 3.12. In both SV and VO phrases, the $S \rightarrow V$ and $O \rightarrow V$ causal fractions were higher whenever the verb was polysemous. The Spearman coefficients and p -values were $\rho = 0.17$, $p < 0.2\%$ for SV phrases and $\rho = 0.16$, $p < 0.4\%$ for VO phrases. In addition, the causal fraction associated with $O \rightarrow V$ was higher whenever objects were

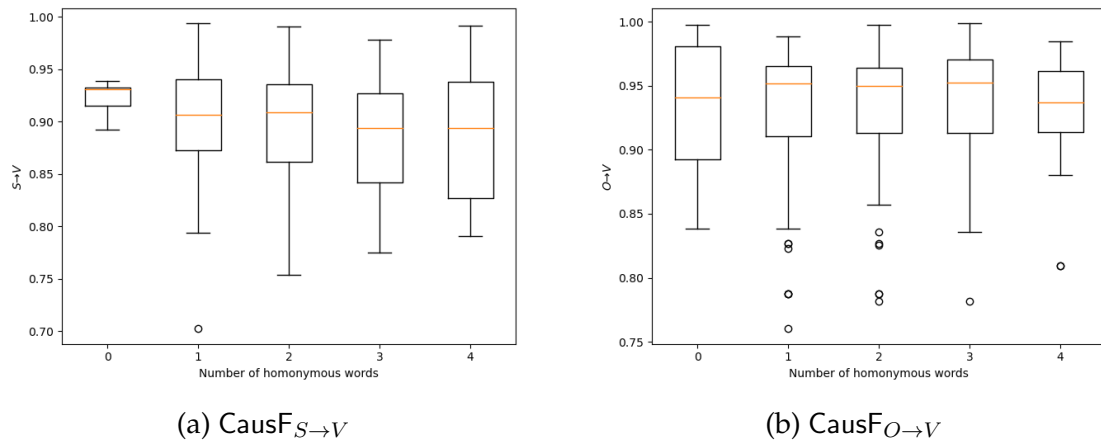
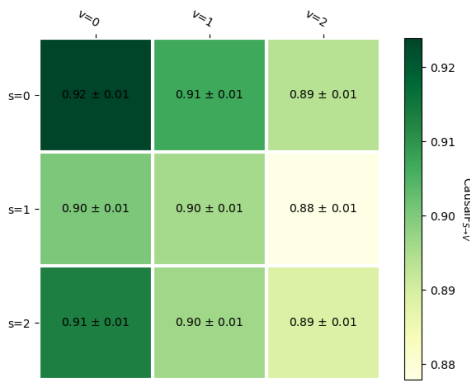


Figure 3.11: Boxplots of the distributions of causal fractions $\text{CausF}_{S \rightarrow V}$ and $\text{CausF}_{O \rightarrow V}$ as a function of the number of homonymous words of the empirical models.

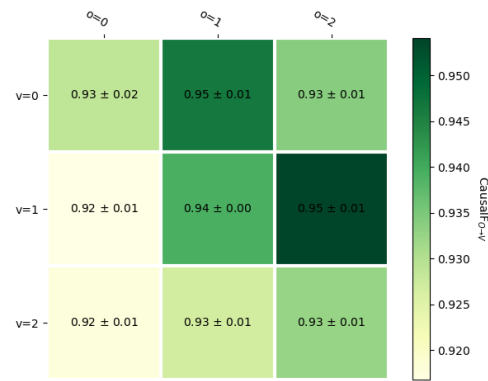
homonymous ($\rho = 0.14$, $p < 2\%$). These observations confirm our initial hypothesis that homonymous nouns are disambiguated faster than polysemous words.

The exception is the case of the ambiguity of the subjects in SV phrases ($\rho = 0.04$, $p > 50\%$). One way to interpret this difference would be to consider the relative position of the disambiguating context. It was shown in [69] that homonymous nouns were disambiguated much faster than polysemous nouns when the disambiguating context occurred before the target words. However, a significant slowdown has been observed if the disambiguating context is found after the target word. This slowdown was even exacerbated when the target word was homonymous. This nicely explains the difference between VO and SV phrases. Indeed, the only possible disambiguation context for nouns in VO phrases is the verb, which is positioned *before* the object (once again assuming active voice). In the case of the disambiguating context for subjects, the only disambiguating context, which is once again the verb, is found after the subject. Therefore, the lack of correlation between the subjects' ambiguity and the causal fraction is likely caused by the balancing of the effect of the ambiguity and the added difficulty induced by a following disambiguating context.

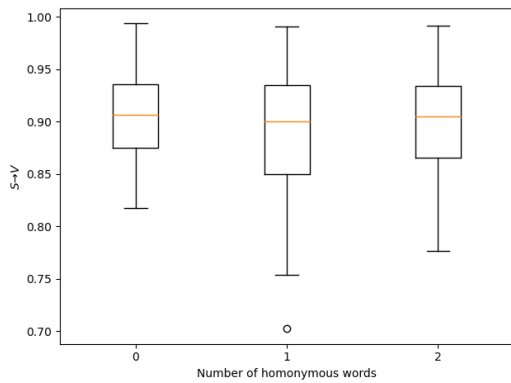
The Spearman coefficients found above are relatively low ($\rho < 0.2$), which suggests that the correlations observed are quite mild. However, the p -values showed that the correlations claimed in the above paragraph are statistically significant, i.e. it is highly unlikely that no correlation exists between the causal fraction and levels of ambiguity.



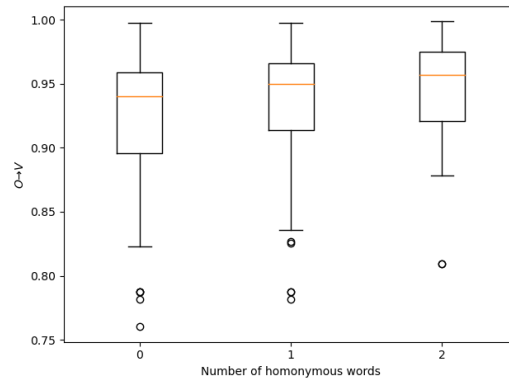
(a) Averaged $\text{CausF}_{S \rightarrow V}$ as the number of homonymous subjects and verbs are varied.



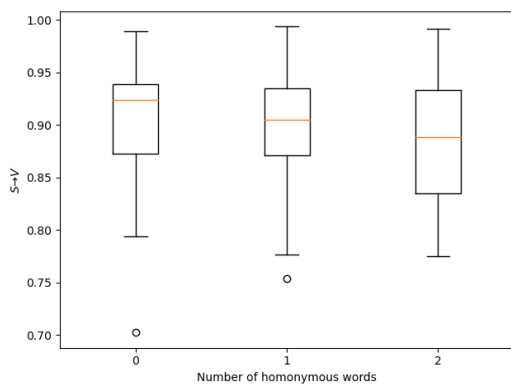
(b) Averaged $\text{CausF}_{O \rightarrow V}$ as the number of homonymous verbs and objects are varied.



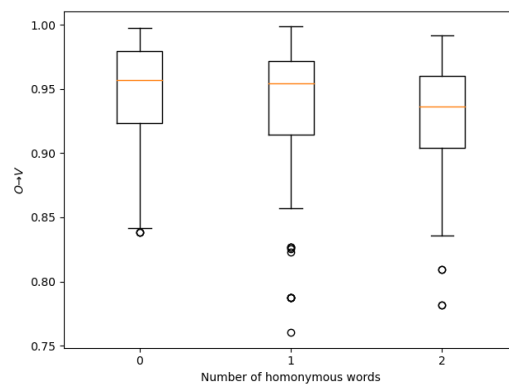
(c) Distributions of $\text{CausF}_{S \rightarrow V}$ as the number of homonymous subjects is varied.



(d) Distributions of $\text{CausF}_{O \rightarrow V}$ as the number of homonymous object is varied.



(e) Distributions of $\text{CausF}_{S \rightarrow V}$ as the number of homonymous verb is varied.



(f) Distributions of $\text{CausF}_{O \rightarrow V}$ as the number of homonymous verb is varied.

Figure 3.12: Causal fractions as the number of homonymous/polysemous nouns and verbs are varied

Summary of the Chapter

In this Chapter, we have investigated the properties of lexical ambiguity data, using the mathematics arising from the causality and contextuality quantum mechanics. We have observed:

- Contextuality-by-Default witnesses can be observed in cyclic systems of rank 2 (see Section 3.2 and [194]). However, the operational interpretation of CbD is not very clear;
- Causal analysis of the data confirms that verbs are mostly disambiguated after their subject and object (see Section 3.4)

Chapter 4

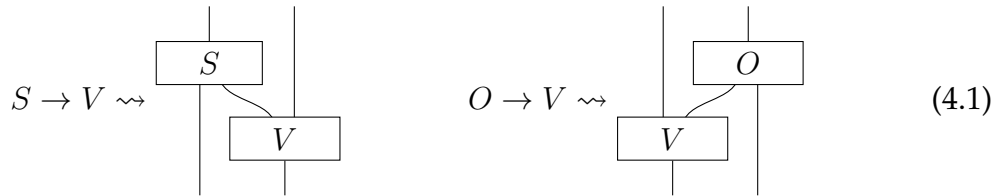
QUANTUM SIMULATIONS OF THE DISAMBIGUATION PROCESS

In the previous chapter, we found that about 90% of the probability distributions obtained from human judgments are compatible with the $S \rightarrow V$ and $O \rightarrow V$ causal orders. Since this dataset is noisy and uses finite approximations of actual probabilities, we can be confident that these causal orders represent a good approximation of the process that occurs in humans. Here, we start from this observation and investigate whether quantum computers can simulate the disambiguation process of SV and VO phrases.

In Section 4.1 we describe how we obtained quantum circuits associated with the lexical disambiguation process of SV and VO phrases. In Section 4.2, we investigate whether the obtained circuits can predict probability distribution instead of reproducing them. In Section 4.3, we describe a method for obtaining quantum word-embeddings of ambiguous words. This section also aims to examine whether we could use the obtained embeddings in NLP. Finally, in Section 4.4 we investigate the entanglement generated by the obtained circuits.

4.1 Methodology

We start from the causal orders obtained in the previous section, i.e. $S \rightarrow V$ and $O \rightarrow V$ for SV and VO phrases. Then, these causal orders naturally lead to a basic structure of the process which we need to approximate (see Section 1.3.3), namely:



These processes should be interpreted as before, i.e. the choice of inputs corresponds to a choice of word, and the outcomes will represent interpretations of these words. In addition, the parties are also defined as in quantum scenarios as “labs” which are allowed to do some local operations (e.g. S , V or O) on the system, and causality, for example, $S \rightarrow V$, is achieved by having a subsystem of the party S being used by the party V .

Remark 4.1. The diagrams of (4.1) are agnostic to which process theory they live on. In particular, we choose quantum circuits in this work, but these would also be applicable in a classical or probabilistic setting as well. Experiments involving classical systems, and hence investigation any potential quantum advantage, is left to future work.

Using this basic structure, we propose a parametric quantum circuit where we can train parameters to approximate the probability distributions obtained from human data. The details of the ansatz of the parametric circuits will be described shortly. We then optimise the parameters of our ansatz using a fairly standard hybrid quantum-classical method [122, 61, 137, 160, 204].

4.1.1 The ansatz

We start by describing the choice of ansatz. For simplicity, we take the input and output systems to be qubits. This will be enough as we only require two choices of inputs, which will be taken to be $|0\rangle$ and $|1\rangle$, and two choices of outcomes, which we will once again choose to be $|0\rangle$ and $|1\rangle$.

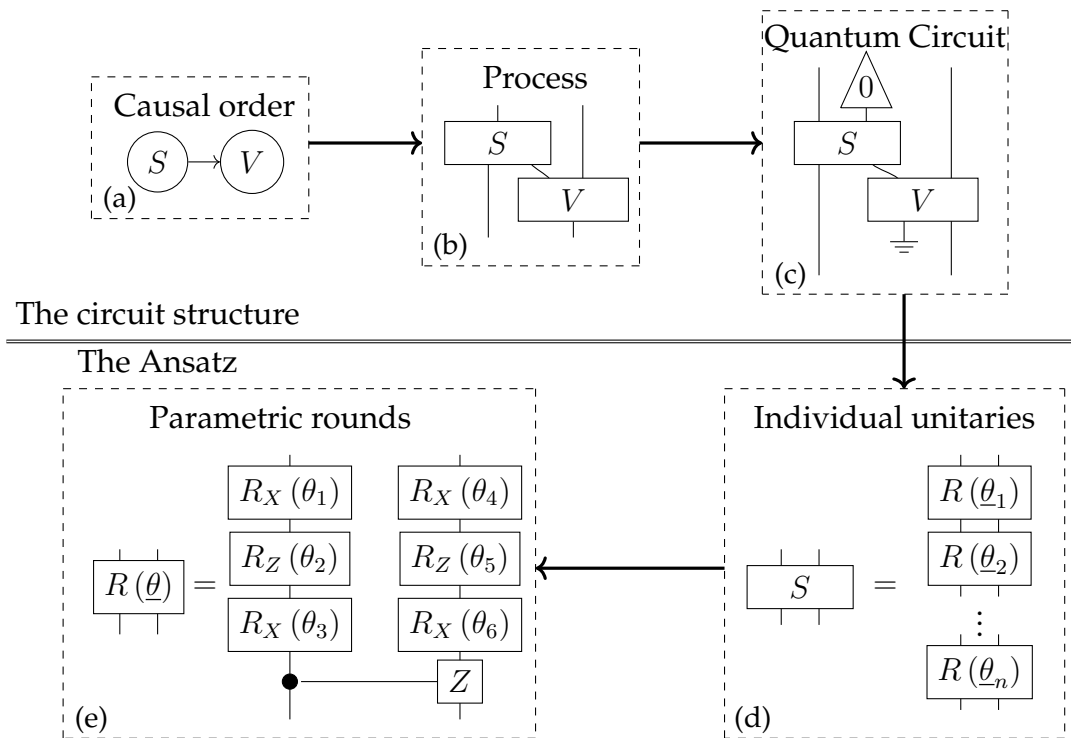


Figure 4.1: Summary of the approach

In addition, we require a subsystem of the output of the noun (subject or object) operations to be fed into the input of the verb operation. We will also take this subsystem to live on \mathbb{C}^2 . Then, to satisfy the unitary condition on the different operations, we will also require that the input system of the noun circuit is a 2-qubit system, where we choose to initialise the ancilla qubit as $|0\rangle$. Similarly, there will be an extra qubit in the output of the verb-circuit, which we will discard. The form of these circuits is illustrated in Fig. 4.1(c).

Then, each of the individual operations will be encoded as a parametric quantum circuit itself. These circuits will be divided into rounds of single-qubit unitaries followed by entangling gates; see Fig. 4.1(d-e). Increasing the number of rounds (and therefore the number of parameters) is expected to increase the accuracy of the circuit but is also expected to take longer to be trained. We will also choose to have the same number of rounds for both parties.

We then choose to define each of the rounds to be as in Fig. 4.1(e). Each qubit is subject to a (parametrised) X -rotation, a (parametrised) Z -rotation, and then an-

other (parametrised) X -rotation, where X - and Z -rotations are defined as follows:

$$R_X(\theta) = \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & -i \sin\left(\frac{\theta}{2}\right) \\ i \sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{pmatrix} \quad R_Z(\theta) = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\theta} \end{pmatrix}$$

This general form allows us to encode any single-qubit unitary, as this corresponds to Euler's decomposition of a generic unitary. We then apply a controlled- Z gate between the two qubits to generate entanglement, defined as^[1]:

$$\text{Circuit Diagram} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

Each round then needs $2 \times 3 = 6$ parameters to be trained. Hence, for n rounds, each gate (S , V or O) needs $6n$ parameters, giving a total of $12n$ parameters to be trained for the full circuit.

Remark 4.2. In most work using variational quantum circuits, the ansatz is less generic and minimises the number of parameters for each round (for example, the IQP ansatz only has 1 parameter per round). However, we decided on this ansatz as it allowed us to access a wide range of *probability distributions*, and because it did converge with respect to our choice of cost function (more details below).

4.1.2 The training process

To train the parameters, we apply a classical gradient descent algorithm. At each iteration of the algorithm, we update parameters θ as follows:

$$\theta_n \rightarrow \theta_{n+1} = \theta_n - \gamma \nabla L(\theta_n) \quad (4.2)$$

for a given cost function L , and descent parameter γ , which we take to be fixed at 10^{-2} . In addition, as the expression of L may not be known, we employ the following finite approximation:

$$(\nabla L(\theta))_i = \frac{L(\theta + \delta \mathbf{e}_i) - L(\theta - \delta \mathbf{e}_i)}{2\delta} \quad (4.3)$$

^[1]Also see Example 1.48 for the definition of general controlled gates.

where δ is chosen to be 10^{-2} . We obtain the cost expression by simulating the quantum circuits using the Qiskit Aer platform [153].

In this task, we are interested in reproducing the probability distributions obtained from human judgments. We, therefore, adopt as the cost function a distance between the obtained probability distribution (estimated from the counts obtained from Qiskit) and the probability distribution obtained from the human judgment dataset. In this work, we adopted the *total variation* between the human and simulated probability distributions as our cost function. It is defined as:

$$L(\theta) = \frac{1}{2} \max_C \sum_o |e(\theta)_C(o) - e_C(o)| \quad (4.4)$$

Remark 4.3. Another choice of cost function would be the Kullback–Leibler (KL) divergence, denoted $D_{KL}(\mu||\nu)$, which measures the expected surprisal induced from using a probability distribution ν to approximate another distribution μ . It is formally defined as:

$$D_{KL}(\mu||\nu) = \sum_x \mu(x) \log \left(\frac{\mu(x)}{\nu(x)} \right) \quad (4.5)$$

However, this measure is a directional measure, i.e. $D_{KL}(\mu||\nu) \neq D_{KL}(\nu||\mu)$, and therefore not a metric. More importantly, it is not defined whenever there is an outcome x such that $\nu(x) = 0$ but $\mu(x) \neq 0$. For these reasons, we chose to use the total variation instead. We will leave the investigation of the circuits obtained using the KL-divergence to future work.

4.1.3 Convergence

We now look at the performance of the described ansatz subject to the proposed training process. We first note that there are some limits to how close the probability distributions obtained from the described circuits can be to the human ones. We will first specify these constraints before introducing the obtained results.

Limitations of the approach

We recall from Section 3.4.1 that, although very high, the causal fractions associated with $S \rightarrow V$ and $O \rightarrow V$ causal order were not exactly 1. As argued before, this is not necessarily because indefiniteness is necessary, in particular when the

causal fraction approaches 1, but can also be due to the finiteness of the probability distribution. Hence, the probability distributions obtained by the variational circuits *cannot* exactly match the probability distributions from the human judgment dataset. In other words, the minimal cost possible will be strictly greater than 0. On the other hand, we can fix a bound on the achievable cost from the causal fraction of a given model.

Proposition 4.4. *Given an empirical model e with parties A and B , with associated causal fraction CausF (with respect to causal order $A \rightarrow B$), for any empirical model e_{Caus} compatible with the causal order $A \rightarrow B$, we have:*

$$\frac{1 - \text{CausF}}{\text{CausF}} m(e) \leq \frac{1}{2} \max_C \sum_o |e_C(o) - e_{\text{Caus},C}(o)| \quad (4.6)$$

where:

$$m(e) = \min \{ e_{(a,b_1)}|_A(1), e_{(a,b_2)}|_A(0) \} \quad (4.7)$$

and $a \in \{a_1, a_2\}$ such that:

$$\text{CausF} = 1 - |e_{(a,b_1)}|_A(0) - e_{(a,b_2)}|_A(0)|$$

The quantity $\frac{1 - \text{CausF}}{\text{CausF}} m(e)$ will be referred to as the minimal cost of an empirical model e .

The proof of this proposition can be found in Appendix C.5.

Outcome of the training process

We trained variational circuits using the above ansatz for numbers of rounds varying from 1 to 5 and set the initial set of parameters randomly. We found that all of the models converged with respect to the cost function (4.4) (see Fig. 4.2a). In addition, the converged cost appears to get closer and closer to the minimal possible cost, dictated by (4.6), as the number of rounds increases (see Fig. 4.2b). This shows that the accuracy of the quantum circuits does indeed increase as the number of parameters increases.

For the rest of this chapter, we assume that models with more parameters are more accurate and, therefore, more representative of the process under investigation.

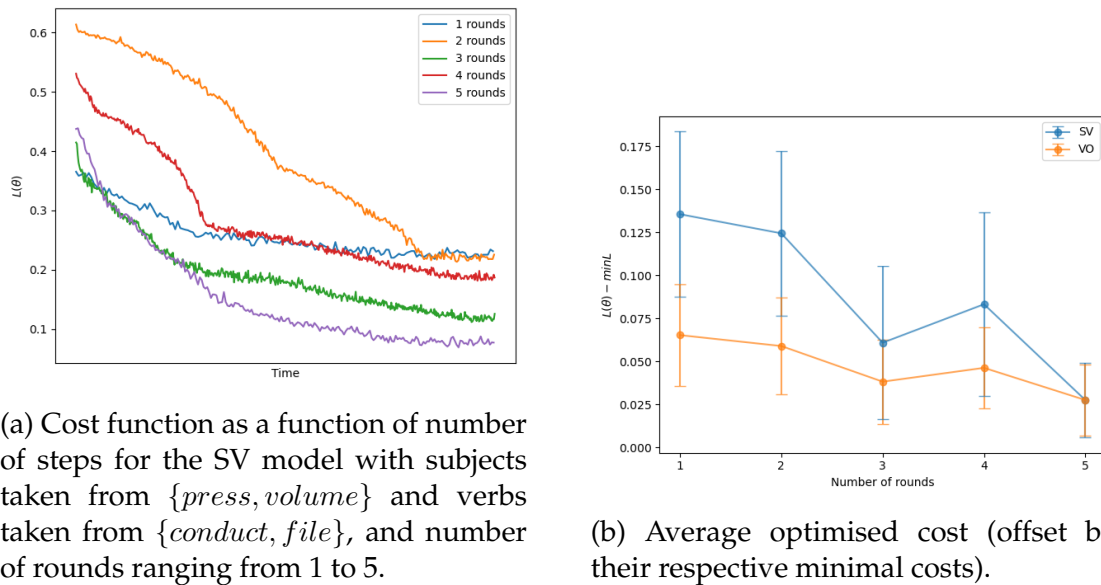
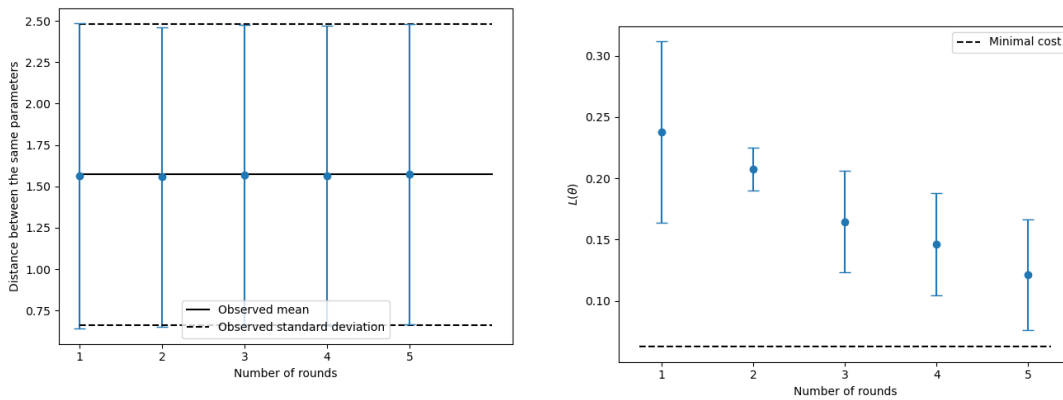


Figure 4.2: Convergence of the variational circuits

On the variability of the optimised parameters We note that, even though the variational circuits converge to similar costs for different choices of initial parameters, the values of the optimised parameters are quite variable for different choices of initial parameters (see Fig. 4.3). The average distance between the parameters for different randomly sampled initial parameters was 1.57 ± 0.91 , which is precisely the expected distance between randomly chosen parameters (which can be calculated to be $\frac{\pi}{2} \pm \frac{\pi}{2\sqrt{3}}$).

This is to some extent expected as, although we know that there exists a minimal possible cost we can achieve, namely (4.6), there is an infinite number of (causal) empirical models e_{Caus} which achieves this minimal cost. This is not necessarily a problem if one is only interested in obtaining (independent) models of the disambiguations of the phrases in a given empirical model. However, in the following discussions, we will be interested in using these circuits for predictions of probability distributions, as well as investigating the use of the obtained quantum states as (quantum) word embeddings.

Hence, from now on, we will fix the choice of initial parameters (for each number of rounds) to reduce the final parameters' variability.



(a) Distances between optimised parameters for different choices of initial parameters.

(b) Optimised cost functions for different choices of initial parameters.

Figure 4.3: Comparison of the accuracies and obtained parameters for the SV model $\{press, volume\} \times \{conduct, file\}$ where the training was done using different initial parameters.

4.2 The prediction power of the variational circuits

We now investigate whether the obtained circuits can predict the activation pattern of the meanings of *unseen phrases*. We then suggest the task of predicting the probability distribution of the different activation patterns for phrases that have not been explicitly trained.

4.2.1 Methods

Here, we propose to obtain predictions by splitting individual operations (i.e. subject, verb, and object) from trained circuits and combining them with operations from different models (see Fig. 4.4). For example, given two optimised SV circuits corresponding to the empirical models with measurements $\mathcal{M}_1 = \{paper, plant\} \times \{bore, tap\}$ and $\mathcal{M}_2 = \{press, volume\} \times \{conduct, file\}$, we can create a new circuit which would correspond to an empirical model with the subjects taken from \mathcal{M}_1 , and the verbs taken from \mathcal{M}_2 , i.e. $\mathcal{M}' = \{paper, plant\} \times \{conduct, file\}$.

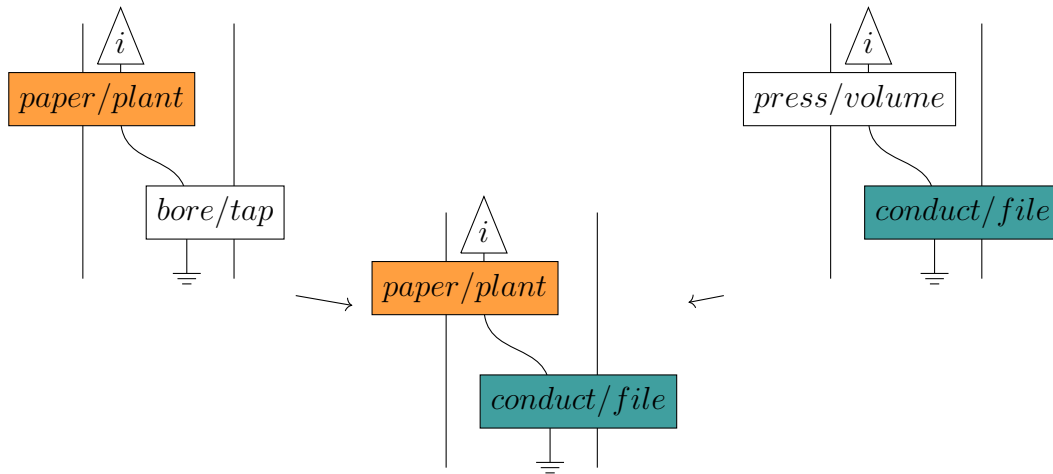


Figure 4.4: Procedure for obtaining new (untrained) circuits from trained ones.

We then constitute a set of 81 SV and 81 VO empirical models, which we will train using the procedure described in Section 4.1. These empirical models will constitute our training set. We then test the predictions obtained from recombining the subject, object, and verb circuits (as in Fig. 4.4) to predict the probability distributions of 84 new SV and 84 VO empirical models for which we have the corresponding human data; these new empirical models will constitute our testing set. The empirical in the training and testing sets can be found in Appendix D.4.

Remark 4.5. Regarding the constitution of the training and testing set, we were restrained by the set of randomly chosen phrases for which we collected the human plausibility judgments. Hence, not all of the possible empirical models arising from the process described above and on Fig. 4.4 could have been evaluated against a ground truth distribution (as we may not have collected it), which in turns restricts our choice of testing set. The training set on the other hand is dictated by which empirical models were needed to obtain predictions for all of the empirical models in the testing set.

4.2.2 Results

We observe that the predicted circuits achieve a reasonably low cost (see Fig. 4.5). The average cost of the unseen models was $\langle L(\theta) \rangle = 0.24$ (i.e. accuracy of 0.76) for the SV models and $\langle L(\theta) \rangle = 0.14$ (i.e. accuracy of 0.86) for the VO models. This procedure resulted in an average cost of 0.19 (i.e. accuracy of 0.81) for both types of

models. These cost values are unsurprisingly still higher than the cost that can be achieved by training the models themselves, which were respectively 0.07 ± 0.04 for SV and 0.06 ± 0.04 for VO models.

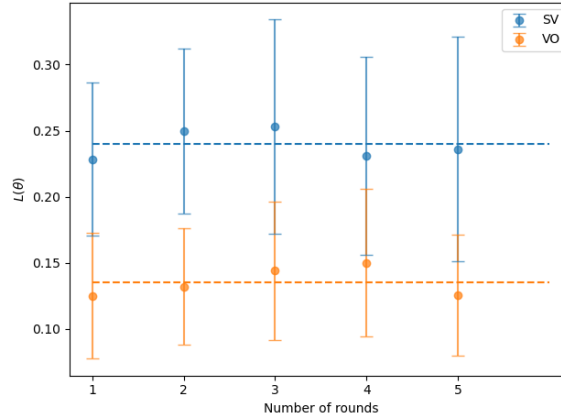


Figure 4.5: Cost values obtained for the predicted probability distributions.

However, these predictions do not improve as the number of rounds increases. Indeed, the Pearson's ρ coefficients between the number rounds and the prediction accuracies are of $|\rho| < 0.01$ (with p -value $p = 0.84$) for the SV models and $|\rho| = 0.06$ and $p = 0.22$ for the VO models; such low values for ρ and higher p -values shows that there likely *no correlation* between the number of rounds and the accuracy of the predictions. On the other hand, the obtained accuracies are significantly better than the ones obtained by taking the uniform probability distributions empirical model (see Table 4.1), which corresponds to the probability distributions obtained by simply guessing outcomes without any knowledge of the system. The differences between the uniform distribution baseline and predicted accuracies were statistically different from 0 with p -values $p < 10^{-10}$ for each number of rounds.

(A, B)	$(0, 0)$	$(0, 1)$	$(1, 0)$	$(1, 1)$
(a_1, b_1)	1/4	1/4	1/4	1/4
(a_1, b_2)	1/4	1/4	1/4	1/4
(a_2, b_1)	1/4	1/4	1/4	1/4
(a_2, b_2)	1/4	1/4	1/4	1/4

Table 4.1: Empirical model containing only uniform probability distributions

4.3 Obtaining quantum word embeddings

Our next step is to see whether the circuits we have trained can be used to give a meaningful representation of words. If this is the case, this would give us a way of obtaining quantum word embedding, which could be used in NLP tasks such as word-sense disambiguation. Testing the performance of such word-state in NLP tasks is beyond the scope of this work and is left to future work.

4.3.1 Methods

Each word-state is trained for a fairly specific empirical model and is not compositional by default. Hence, we first want to check that the states which should correspond to the same word are indeed similar. If this were not the case, these word-states would not be useful anyway as a single word would have multiple representations. Here, we will assume that each word's dependency constitutes an inherent part of the word. For example, the noun *pitcher* used as a subject will be considered distinct from the same word used as an object. Similarly, the verb *tap* taking a subject will be considered distinct from the same verb taking an object instead.

As we did for the prediction task, we first fixed the initial parameters for all of the variational circuits (otherwise, we would expect the different word-states to share little features). Then, using the optimised circuits, we can obtain a quantum state representation of a subject or object by fixing the input of the S or O individual circuits. For example, given the circuit representation of $\{press, volume\}$:

$$\begin{array}{|c|} \hline S \\ \hline \end{array} = \begin{array}{|c|} \hline press/volume \\ \hline \end{array} \quad (4.8)$$

(obtained from, say, training for the empirical model corresponding to the SV model $\{press, volume\} \times \{conduct, file\}$), we can obtain a quantum state representation of

the word *volume* (as a subject) as:

$$\text{triangle}_{\text{volume}_S} = \text{circuit}_{\text{press/volume}} \quad (4.9)$$

(recall that, in the case of SV circuits, the right-hand ancilla state is always set to $|0\rangle$ for SV circuits). Similarly, given the verb-object model associate with $\{\textit{conduct}, \textit{file}\} \times \{\textit{press}, \textit{volume}\}$, given an optimised object circuit:

$$\text{triangle}_{\text{volume}_O} = \text{circuit}_{\text{press/volume}} \quad (4.10)$$

the quantum state representation of the word *volume* (as an object) is:

$$\text{triangle}_{\text{volume}_O} = \text{circuit}_{\text{press/volume}} \quad (4.11)$$

(recall that, in the case of VO circuits, the right-hand ancilla state is always set to $|0\rangle$). Also note that the RHS of (4.8) and (4.10) do not have to be related, so in general, (4.9) will be different from (4.11).

For verbs, the process is a little more complicated as the representation of the verb not only depends on the choice of the verb but is also taking some information from the *S* or *O* circuit as well (see Fig. 4.1). Hence, to obtain the verb representation, we first fix its input and then take the *partial trace* over the subsystem dependent on the *S* or *O* output. This procedure will give us a density matrix instead of a pure quantum state. For example, given the verb circuit form the SV circuits obtained for the empirical model $\{\textit{press}, \textit{volume}\} \times \{\textit{conduct}, \textit{file}\}$:

$$\text{triangle}_{\text{volume}_O} = \text{circuit}_{\text{press/volume}} \quad (4.12)$$

we can obtain the representation of the verb *conduct* as:

$$(4.13)$$

In order to quantify “how similar” two word-vectors are, we decide to calculate the inner products between them. Here, we will be interested in the inner-product between states representing the same word (and dependency) but have been trained to approximate different empirical models. For example, given two SV circuits for the empirical models associated with $\mathcal{M}_1 = \{press, volume\} \times \{conduct, file\}$ and $\mathcal{M}_2 = \{line, volume\} \times \{box, reflect\}$, we would like to compare the word-states associated with *volume* in the two optimised circuits.

In the case of nouns (i.e. subjects or objects), we recall that these are pure quantum states, so we calculate their inner product using the Born rule:

$$|\langle \psi_n | \phi_n \rangle|^2 = \left| \begin{array}{c} \triangle \phi_n \\ | \\ \triangle \psi_n \end{array} \right|^2 \quad (4.14)$$

where $|\psi_n\rangle$ and $|\phi_n\rangle$ are both representation of the noun n . For example, given the circuits optimised for the empirical models associated with \mathcal{M}_1 and \mathcal{M}_2 as defined above in the paragraph, we define the inner products of the two word-states associated with the word *volume* as:

$$(4.15)$$

For verbs, which we recall are density matrices, we apply the generalised Born

rule instead to calculate the inner product:

$$\text{Tr}(\rho_v \varrho_v) = \begin{array}{c} \triangle \rho_v \\ | \\ \triangle \varrho_v \end{array} \quad (4.16)$$

where ρ_v and ϱ_v are both representation of a verb v . For example, given two SV circuits optimised to approximate the empirical models associated with :

$$\begin{aligned} \mathcal{M}_1 &= \{press, volume\} \times \{conduct, file\} \\ \mathcal{M}_3 &= \{letter, paper\} \times \{conduct, grasp\} \end{aligned}$$

we can define the inner-product between the two word-states of the verb *conduct* as:

In both cases, an inner-product of 1 will mean that the two states are equivalent, and an inner-product of 0 will mean that the two states share no feature.

4.3.2 Results

We found out that the pure states corresponding to nouns have a larger overlap between different models, where the average inner product was of 0.64 for noun-states, compared to mixed verb-states where the average inner product was 0.37.

In addition, we observed stark differences between these inner products depending on whether the word states were associated within the same or different input state, where we only consider the input state which identifies the word of interest (as opposed to the ancilla input, which is always the same). For example, the word *press* in the model $(press, volume) \times (conduct, file)$ is associated with the input state

$|0\rangle$, whereas the same word in the model $(line, press) \times (admit, wipe)$ will correspond to the input state $|1\rangle$.

In particular, we observed that the word states associated with words corresponding to the same input states (e.g. *press* in $(press, volume) \times (conduct, file)$ and $(press, television) \times (box, label)$) have an inner product close to 1. The average of these inner-product was 0.94 for pure noun-states and 0.48 for mixed verb-states.

By contrast, the word states corresponding to different input states (e.g. *press* in $(press, volume) \times (conduct, file)$ and $(line, press) \times (admit, wipe)$) had a small overlap (on average the inner-product between them was 0.03 for noun states and 0.02 for verb states).

This is expected as we have previously seen that the output of the training process is highly dependent on the choice of initial parameters. Hence, when words are associated with different input states, they will not necessarily correspond to the same optimised circuits.

Overall, this means that the quantum word embeddings that we obtain from these variational circuits have the potential to be useful in NLP tasks as long as one fixes the input state it corresponds to.

4.4 Entanglement of phrases and words

We now investigate how much entanglement has been created using the optimised variational circuits. Entanglement is often considered the primary source of quantum correlation [150, 188]. A quantum state is said to be *entangled* (or *non-separable*) iff it cannot be prepared using Local Operations and Classical Correlations (LOCC) alone [144]. If this amount of entanglement is high, in particular for the more accurate models, this suggests that training using quantum resources may be beneficial (as opposed to simply using classical probabilistic methods).

4.4.1 Entanglement measures

Due to its importance in quantum information theory, entanglement needs to be quantified. Many measures of entanglement have been proposed for bipartite and multipartite states, for pure and mixed states. Here, we will focus on the bipartite

measures and introduce an entanglement measure pure state and one for mixed states.

For pure state, the standard measure of entanglement is the *entanglement entropy*, defined as:

$$E(|\psi\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B) = -Tr(\rho_A \log_2 \rho_A) = -Tr(\rho_B \log_2 \rho_B) \quad (4.18)$$

where $\rho_{A,B} = Tr_{A,B} |\psi\rangle \langle \psi|$.

Many (non-equivalent) measures of entanglement have been proposed for mixed states. Among which is the *entanglement of formation*, formally defined as:

$$E_F(\rho) = \inf \left\{ \sum_k p_k E(|\psi_k\rangle) \mid \rho = \sum_k p_k |\psi_k\rangle \langle \psi_k| \right\} \quad (4.19)$$

Using the above definition alone, it is very hard to calculate the entanglement of formation for an arbitrary density matrix, as it involves finding all of the possible decompositions of the matrix ρ in terms of density matrices of pure states. Fortunately, in the case of qubit systems, a closed formula has been found to calculate the entanglement of formation [199], namely:

$$E_F(\rho) = s \left(\frac{1 + \sqrt{1 - C^2(\rho)}}{2} \right) \quad (4.20)$$

where $C(\rho)$ is defined as:

$$C(\rho) = \max \{0, \lambda_1 - \lambda_2 - \lambda_3 - \lambda_4\} \quad (4.21)$$

where $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ is the ordered set of eigenvalues of $\rho \sigma_Y \otimes \sigma_Y \rho^* \sigma_Y \otimes \sigma_Y$, and s is defined as:

$$s(x) = -(x \log_2 x) - ((1-x) \log_2(1-x)) \quad (4.22)$$

4.4.2 Entanglement of the optimised circuits

We start by looking at the amount of entanglement created by the whole parametric circuits.

These parametric circuits are bipartite by design, and only make use of qubits. However, we discard a qubit system in each of the circuits. Hence, we will need to quantify entanglement for mixed states.

Results & Discussion The amount of entanglement of formation for the trained parametric circuits is depicted in Fig. 4.6. As we can see, the amount of quantum correlations seems to increase as the circuit’s accuracy increases, particularly for VO models.

This would suggest that the process of disambiguation is “truly parallel” instead of having probabilistic mixtures of combinations of interpretations that can be selected, particularly in VO phrases.

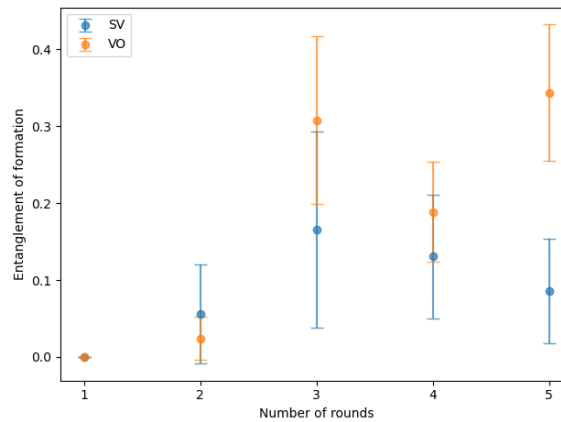


Figure 4.6: Entanglement of formation of the optimised parametric circuits, as the number of rounds increase.

4.4.3 Entanglement of the noun-embeddings

We now want to study the amount of entanglement created for each word individually. However, since only one qubit in the output of verb-circuits is not discarded, any meaningful representation of verb vectors would be simply monopartite, and the notion of entanglement does not apply. We then primarily focus on the degrees of entanglement of nouns. In addition, since subject and object word-states are pure states by design, we can use entanglement entropy (4.18) to measure the amount of quantum correlations created.

The evolution of the entanglement entropy of noun states as the number of rounds in the ansatz increases is shown in Fig. 4.7. The entanglement entropy of subjects is high for smaller numbers of rounds ($\langle E \rangle = 0.95 \pm 0.03$ for $n = 1$ rounds) but decreases as the number of rounds increases ($\langle E \rangle = 0.25 \pm 0.18$ for $n = 5$ rounds). The opposite happens for objects, that is, the amount of entanglement of noun states increases as the approximations get better and better (from $\langle E \rangle = 0.55 \pm 0.04$ for $n = 1$ round to $\langle E \rangle = 0.83 \pm 0.10$ for $n = 5$ rounds).

This observation suggests that the correlations between the subject and the verb are not as strong as the correlations between the object and the verb, or that SV phrases are disambiguated more locally than VO phrases where more interaction between the two words would be needed. As before, this would require further experiments to confirm this trend.

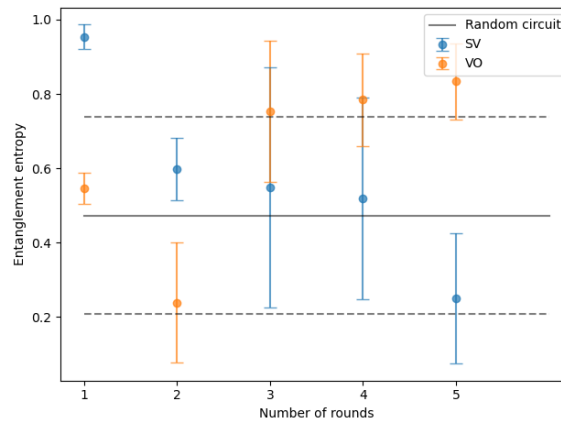


Figure 4.7: Evolution of the entanglement generated by the subject and object circuits, as the number of rounds increases.

In addition, the entanglement entropy of the nouns (subject or object) does not appear to depend on whether it is homonymous or polysemous. All of the p -values were $p > 0.05$ for subjects and $p > 0.19$ for objects.

However, this is by design, as nouns are represented by pure states. Hence, the only way to obtain correlations between the nouns and the verb is through entanglement. It would be interesting to train density matrices in future work and check whether the ambiguity of the nouns affects the entanglement of its quantum state representation.

Summary of the Chapter

- Using the observations of Chapter 3, we proposed a quantum model of the disambiguation of subject-verb and verb-object phrases.
- This model was successfully implemented by variational circuits.
- The optimised quantum circuits were used to predict the meaning of unseen phrases.
- Preliminary evidence suggest that these circuits could be used as embeddings in NLP tasks.
- The optimised circuits generated a non-negligible amount of entanglement.

Part III

SYNTACTIC AMBIGUITY

Chapter 5

MODELING THE HUMAN PARSING PROCESS

In this chapter, we introduce our models of the human parsing process using sheaf theory. We adopt a model in which all possible parses are available at any given stage of a sentence with some probabilities (which we estimate empirically in Chapter 6). This description is compatible with parallel-ranked processing, where the reader constructs parses in parallel and associates each parse with a weight, and probabilistic serial strategies, where the reader creates each parse with a given probability.

This model leads to empirical models similar to the ones described in the previous part and the ones of quantum mechanics [6, 79, 80, 5]. Our empirical models only consider syntactic parses and no lexical or discourse information. Therefore, this is a preliminary model, and our results suggest that such models can indeed be used to represent human processes.

In Section 5.1, we present the structure of our models and how empirical statistics are collected. In Section 5.2, we formally analyse the empirical models regarding quantum contextuality and causality. This section also includes intuitions about interpreting the signalling and causality in the models.

These models, as well as the predictions described in the next chapter can also be found in [193].

5.1 A sheaf-theoretic model of the syntax of sentences

We start by describing a sheaf-theoretic model of human parsing based on the following key points:

1. **Incrementality.** We want to study the evolution of the reader's mental representation as they encounter more information. The literature shows that human understanding is highly incremental. Therefore, we want to create a model that follows the linear order of the words in the sentence, i.e. how information is presented to the reader.
2. **Grammatical structure.** This is our main object of interest. We want to capture what the reader thinks of the grammatical structure of a sentence or part of a sentence. In this work, only the grammatical structure is taken into account. In future work, we plan to include other factors such as plausibility, thematics, etc.
3. **Statistics.** We follow the psycholinguistic hypothesis that the reader may keep in mind *all* of the possible grammatical structures at each stage, but with different ratings (see Section 2.2.1). This implies that we need a way of "rating" different grammatical structures. In the following sections, we opt for probabilities over partial parses. This will give a parallel-ranked model of the parsing process.

Sheaves are a promising way of combining these concepts in a single framework. The intuition is that the statistics are defined over the set of possible grammatical structures. In turn, the grammatical structures are only defined over a set of words that are presented to the reader, and evolve in a linear fashion. Let us describe this idea in more detail.

5.1.1 Incrementality

As described above, we take our contexts to be words as appearing in a sentence or a phrase. The combination of these words forms sentence fragments by concatenation. Now, for a given sentence, e.g. *The employees understood the contract*, we can define many different sentence fragments, for example, *The employees*, *The employees understood*, *employees understood*, etc. We define a *prefix order* over this set of sentence fragments as follows:

Definition 5.1. A fragment s_1 of a sentence is included in another fragment s_2 iff s_1 is a prefix of s_2 . We write:

$$s_1 \leq_p s_2 \quad (5.1)$$

The set of fragments of a sentence S equipped with the prefix order forms a (pre-order) category \mathcal{C}_S .

Example 5.2. In the sentence *The employees understood the contract*, we have:

$$\textit{The employees} \leq_p \textit{The employees understood} \quad (5.2)$$

However, the two fragments *The employees* and *employees understood* are not comparable.

Remark 5.3. We could have chosen morphism to be the simple inclusion of sub-phrases, e.g. taking: $\textit{employees} \subseteq \textit{The employees}$. However, in the case of a purely incremental model of parsing, the prefix order appears to be the most relevant, as it models the order of information available to the reader.

In order to study the incremental evolution of the probability distributions, we consider a *sequence* of empirical models. Each empirical model will consist of two *consecutive stages*. For example, if we want to study the behaviour at the word level (e.g. the difficulty of reading a word), we would take stages to be words and subsequently consider sequences of empirical models containing a pair of contexts, where the contexts only differ by one word. In this case, the sentence *The employees understood the contract would change* would lead to the sequence of empirical models with

contexts:

$$\begin{aligned}
 \mathcal{M}_1 &= \{The, The\ employees\} \\
 \mathcal{M}_2 &= \{The\ employees, The\ employees\ understood\} \\
 \mathcal{M}_3 &= \{The\ employees\ understood, The\ employees\ understood\ the\} \\
 &\vdots \\
 \mathcal{M}_6 &= \{The\ employees\ understood\ the\ contract\ would, \\
 &\quad The\ employees\ understood\ the\ contract\ would\ change\}
 \end{aligned}$$

Similarly, we could take stages to correspond to regions or phrases of the sentence, in which case we would consider sequences of empirical models in which the contexts differ by a region. An example of such a sequence of empirical models would consider the following contexts:

$$\begin{aligned}
 \mathcal{M}_1 &= \{The\ employees, The\ employees\ understood\ the\ contract\} \\
 \mathcal{M}_2 &= \{The\ employees\ understood\ the\ contract, \\
 &\quad The\ employees\ understood\ the\ contract\ would\ change\}
 \end{aligned}$$

We could also consider a more fine-grained analysis and take tokens or morphemes to be the incremental unit. In this chapter, we will study word-by-word parsing behaviour.

5.1.2 Grammatical structures

For each phrase or context, we want to be able to associate a grammatical structure. The grammatical structure will then be the *outcomes* of our models. There are different ways to represent the grammatical structure of natural language input, including constituency trees [39], dependency grammars [158], and categorial grammars [11, 19, 112, 170]. In this work, we decide to work with dependency grammars.

Dependency grammar

In dependency grammar, each word of the sentence is associated with a *head* and a *dependency* or syntactic function. The main dependency structures are as follows:

- The main verb of a sentence, or the main noun of a noun-phrase, is its own head and is associated with the dependency `ROOT`;
- The head of the subject of a verb is the verb, and its dependency is `nsubj` (nominal subject);
- The head of the object of a transitive verb is once again the verb, and its dependency is `dobj` (direct object);
- The head of a determiner is its head noun and comes with dependency `det`;
- The head of an adjective is the noun it is modifying, and its dependency is `amod` (adjective modifier);
- ...

These dependency structures are usually represented as graphs, where we use labelled directed arrows $word \xrightarrow{dependency} head(word)$ to represent them (see Fig. 5.1 for example of such graphs).

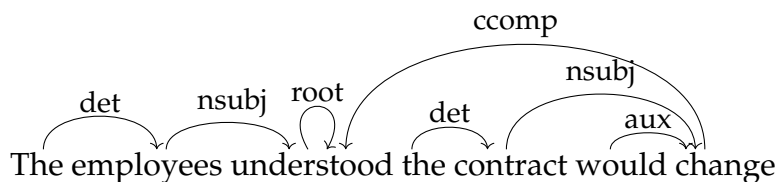


Figure 5.1: Dependency relations in the sentence *The employees understood the contract would change*.

The presheaf of events

Events The main reason for adopting dependency grammar instead of a different paradigm is its minimality. In particular, it is easy to convert a dependency graph to a function $s : U \rightarrow \mathbb{N} \times D$, where U corresponds to the *ordered* set of words in a sentence, and D is the set of dependency relations.

To do so, we start by labelling the words of the sentence (or sentence fragment) by their position in the sentence (resp. sentence fragment) – then each word has a label in \mathbb{N} . Then, each element of U can be seen as an element of $\mathcal{V} \times \mathbb{N}$, where \mathcal{V} is the vocabulary. We want a pair $(w, n) \in \mathcal{V} \times \mathbb{N}$ to represent a word w being found in position n in a sentence fragment. Hence, we would also require that for each element of U :

$$(w, k), (w', k) \in U \implies w = w' \quad (5.3)$$

This last condition states that we can only have one word at position k . We can then see that the objects of \mathcal{C}_S , can be expressed in this fashion as:

$$w_1 \dots w_n := \{(w_1, 1), \dots, (w_n, n)\} \quad (5.4)$$

Example 5.4. Consider the sentence $S = \text{The employees understood the contract would change}$. Here, we have:

$$\begin{aligned} \text{The} &= \{(The, 1)\} \\ \text{The employees} &= \{(The, 1), (employees, 2)\} \\ \text{The employees understood} &= \{(The, 1), (employees, 2), (understood, 3)\} \\ &\vdots \\ S &= \{(The, 1), (employees, 2), (understood, 3), (the, 4), \\ &\quad (contract, 5), (would, 5), (change, 6)\} \end{aligned}$$

We are now ready to define the dependency structure of a sentence fragment U as a function $s : U \rightarrow \mathbb{N} \times D$, such that $(w, k) \mapsto (k', d)$ signifies that the word w (at position k) has a head at position k' with dependency d . So, for instance:

$$\begin{array}{c} \text{det} \quad \text{nsbj} \quad \text{root} \\ \text{The} \quad \text{employees} \quad \text{understood} \end{array} \equiv \begin{cases} (The, 1) & \mapsto (2, \text{det}) \\ (employees, 2) & \mapsto (3, \text{nsbj}) \\ (understood, 3) & \mapsto (3, \text{ROOT}) \end{cases} \quad (5.5)$$

Presheaf structure Using the correspondance (5.4) between the defined sets U and objects of \mathcal{C}_S , we can define the functor:

$$\begin{aligned} \tilde{\mathcal{E}} : \quad \mathcal{C}_S^{op} &\rightarrow \mathbf{Sets} \\ U &\mapsto \{\tilde{s} : U \rightarrow \mathbb{N} \times D\} \\ U \leq_p V &\mapsto \tilde{s}_V \mapsto \tilde{s}_V|_U \end{aligned} \quad (5.6)$$

where the restriction morphisms are defined $s_V \mapsto s_V|_U$ as:

$$\tilde{s}_V|_U((w, k) \in U) = \tilde{s}_V((w, k) \in V) \quad (5.7)$$

Note that this is well-defined since $U \leq_p V$ implies that U and V are of the form:

$$\begin{aligned} U &= \{(w_1, 1) \dots (w_n, n)\} \\ V &= \{(w_1, 1) \dots (w_k, k)\} \end{aligned}$$

where $n \leq k$.

Example 5.5. Let us again take the sentence $S =$ *The employees understood the contract would change.* For $U =$ *The employees* and $V =$ *The employees understood*, we would have:

$$\begin{array}{c} \begin{array}{ccc} \text{det} & \text{nsubj} & \text{root} \\ \text{The employees understood} \end{array} \\ \left| \begin{array}{c} \text{The employees} \end{array} \right. \\ \end{array} = \begin{array}{ccc} \text{det} & \text{nsubj} & \\ \text{The employees} & [\dots] & \end{array} \quad (5.8)$$

Remark 5.6. Since we are dealing with sentence fragments as well as sentences, the head of a word in a fragment may be *undefined* or at least *underspecified*. This feature is quite useful from a cognitive plausibility point of view, as it is hypothesised that humans tend to make predictions about the completion of sentences. Therefore, dependencies may not be known in advance.

To simplify calculations, we restrict our data by only considering unlabelled attachments, i.e., to only consider the head of each word as a grammatical structure. One can see that this is enough to distinguish between the different syntactic structures of NP/S and NP/Z garden-path sentences (see Fig. 5.2). Hence, we will take

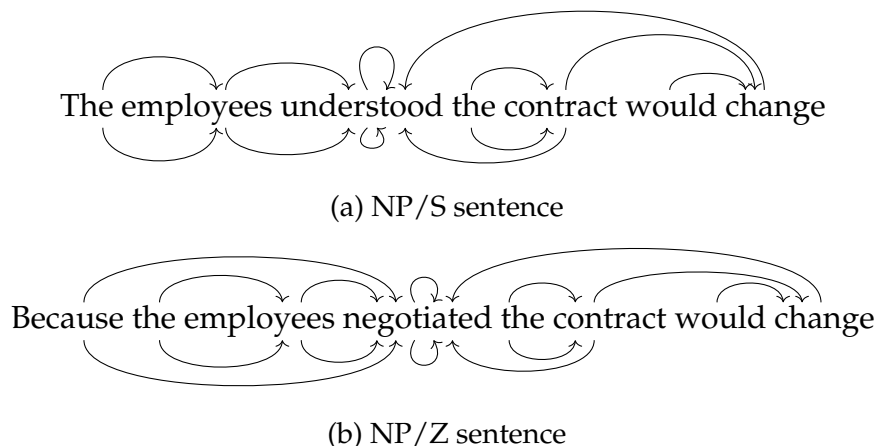


Figure 5.2: Examples of unlabelled dependency parses for NP/S and NP/Z sentences. The discarded parse is also shown (bottom/upside-down parse).

our *presheaf of events* to be the functor:

$$\begin{aligned}
 \mathcal{E} : \quad \mathcal{C}_s^{op} &\rightarrow \text{Sets} \\
 U &\mapsto \{s : U \rightarrow \mathbb{N}\} \\
 U \leq_p V &\mapsto s_V \mapsto s_V|_U
 \end{aligned}
 \tag{5.9}$$

where the parses $s : U \rightarrow \mathbb{N} \in \mathcal{E}(U)$ can be obtained from the parses $\tilde{s} : U \rightarrow \mathbb{N} \times D \in \tilde{\mathcal{E}}(U)$ as:

$$s = \pi_1 \circ \tilde{s} \tag{5.10}$$

Example 5.7. Taking a labelled parse to be:

$$\tilde{s} = \text{The} \overset{\text{det}}{\curvearrowright} \text{employees} \overset{\text{nsubj}}{\curvearrowright} [\dots]$$

the associated unlabelled parse is given as:

$$s = \text{The} \overset{\curvearrowright}{\curvearrowright} \text{employees} \overset{\curvearrowright}{\curvearrowright} [\dots]$$

5.1.3 Probability distributions

We now associate probability scores with possible parses. As in previous chapters, this is done by selecting sections of the presheaf of events post-composed with the

distribution monad $\mathcal{D}_{\mathbb{R}_+} : \text{Sets} \rightarrow \text{Sets}$.

We recall that the functor $\mathcal{D}_{\mathbb{R}_+}$ associate with each set A , the set of probability distributions over A . Hence, for each $U \in \text{ob}(\mathcal{C}_S)$, the set $\mathcal{D}_{\mathbb{R}_+}\mathcal{E}(U)$ corresponds to the set of all probability distributions over the set of possible (dependency) parses of U . Now, we are only interested in *one* probability distribution in the set $\mathcal{D}_{\mathbb{R}_+}\mathcal{E}(U)$, namely, the probability distribution over parses corresponding the mental representation of the syntactic structure of the fragment U .

Example 5.8. For $U = \text{The employees understood}$, we could single out the following probability distribution $e_{\text{The employees understood}} \in \mathcal{D}_{\mathbb{R}_+}\mathcal{E}(U)$:

$$\begin{aligned}
 e_{\text{The employees understood}} \left(\begin{array}{c} \text{The employees understood } [\dots] [\dots] [\dots] [\dots] \\ \text{The employees understood } [\dots] [\dots] [\dots] [\dots] \end{array} \right) &= 0.95 \\
 e_{\text{The employees understood}} \left(\begin{array}{c} \text{The employees understood } [\dots] [\dots] [\dots] [\dots] \\ \text{The employees understood } [\dots] [\dots] [\dots] [\dots] \end{array} \right) &= 0.02 \\
 e_{\text{The employees understood}} (\text{other syntactic structures}) &< 0.01
 \end{aligned} \tag{5.11}$$

Recall that we consider a sequence of empirical models such that in each empirical model, the measurement scenario consists of a pair of contexts differing by a single word. Consequently, an empirical model will consist of a pair of probability distributions.

Example 5.9. An empirical model corresponding to \mathcal{M}_3 of the sentence *The employees understood the contract would change* could consist of the probability distribution of (5.11) and :

$$\begin{aligned}
 e_{\text{The employees understood the}} \left(\begin{array}{c} \text{The employees understood the } [\dots] [\dots] [\dots] \\ \text{The employees understood the } [\dots] [\dots] [\dots] \end{array} \right) &= 0.37 \\
 e_{\text{The employees understood the}} \left(\begin{array}{c} \text{The employees understood the } [\dots] [\dots] [\dots] \\ \text{The employees understood the } [\dots] [\dots] [\dots] \end{array} \right) &= 0.35 \\
 e_{\text{The employees understood the}} \left(\begin{array}{c} \text{The employees understood the } [\dots] [\dots] [\dots] \\ \text{The employees understood the } [\dots] [\dots] [\dots] \end{array} \right) &= 0.26 \\
 e_{\text{The employees understood the}} \left(\begin{array}{c} \text{The employees understood the } [\dots] [\dots] [\dots] \\ \text{The employees understood the } [\dots] [\dots] [\dots] \end{array} \right) &= 0.01 \\
 e_{\text{The employees understood the}} (\text{other syntactic structures}) &< 0.01
 \end{aligned} \tag{5.12}$$

5.2 Contextuality, causality and signalling of the models

We now investigate the properties of the created empirical models. In particular, we are going to focus on the properties of the signalling fraction SF, which was first defined in [183] and Section 1.2.2, and is going to be our main reading time predictor in Chapter 6. First, we start with the impossibility of observing contextuality in this type of scenario.

5.2.1 Contextuality

Not all of the measurement scenarios are capable of hosting contextuality. This can be determined by looking at the structure of \mathcal{M} . In particular, it is known from Vorob'ev's theorem [189, 20] that if the set $\mathcal{M} = \{M_1, \dots, M_n\}$ satisfies^[1]:

$$M_1 \subseteq M_2 \subseteq \dots \subseteq M_n \quad (5.13)$$

then, any compatible family $\{e_M \mid M \in \mathcal{M}\}$ of probability distribution over \mathcal{M} admits a global distribution over $\bigcup_{M \in \mathcal{M}} M$, which can in fact be shown to be the maximal distribution e_{M_n} using the compatibility assumption. In other words, the empirical models described in Section 5.1 cannot exhibit contextuality.

5.2.2 Causality and signalling

Here, we argue that the linguistic models correspond to both a contextuality and a causality scenario. To see this, we first note that for each empirical model, one context includes exactly one less word than the other. As a result, we can without loss of generality see empirical models as an $\{m, mw\}$ scenario. For example, in the empirical model M_2 , we have $m = \textit{The employees}$ and $w = \textit{understood}$. We can, therefore, express the compatibility relation of this model as:

1. A symmetric relation analogous to measurements that can be simultaneously measured. In this case, we interpret m and w as compatible, meaning they are

^[1]Topologically speaking, this means that \mathcal{M} forms a simplex.

somewhat parsed independently. In this interpretation, we have a situation similar to contextuality scenarios;

2. An asymmetric relation analogous to causal scenarios. In this case, the compatibility relation \preceq is read $m \preceq w$.

Both interpretations are possible and very much related. However, the meanings of the quantity SF are subtly different.

In the symmetric interpretation, SF quantifies how consistent the two probability distributions are. On the other hand, in the causal interpretation, the signalling fraction is also a measure of the departure from a causal model following the linear *reading* order. In other words, a high signalling fraction is evidence that parsing a particular subphrase is not incremental^[2] but instead should require information coming from words situated *after* the phrase under consideration.

In either case, if we observe that an empirical model has a higher signalling fraction, this should signify that some reanalysis has to occur. According to psycholinguistic parsing theories, this should trigger a slowdown in reading time.

Hence, we hypothesise that the signalling fraction SF, equivalently the non-causal fraction NCausF, should correlate with human reading times. We investigated this in Section 6.2. For uniformity purposes, we will focus on the models' signaling in the rest of this part.

5.2.3 Computing SF

Computing the signalling/causal fractions in a generic empirical model is not a trivial task, as it requires finding a solution to a linear optimisation problem [183] (see also the discussion in Section 3.4). However, given the specific structure of our empirical models, it is possible to find an expression of the signalling fraction SF, which can be calculated efficiently.

Proposition 5.10. *The signalling fraction can be computed via the following equation*

$$\text{SF} = 1 - \sum_o \min(e_{mw|_m}(o), e_m(o)) \quad (5.14)$$

^[2]In the linguistic sense, i.e. following the left-to-right reading order.

The proof of this proposition can be found in Appendix C.6.

We argue that the signalling fraction measures parsing difficulty. This claim is motivated by the fact that SF can be seen as a measure of distance between probability distributions observed at different stages of the sentence. Therefore, the higher the signalling fraction, the more readers will have to readjust their mental representation of the grammatical structure. We can even say that since the contexts $m_i, m_{i+1} \in \mathcal{M}_i$ only differ by a single word, the signalling fraction of the empirical model e_i becomes related to the difficulty of understanding the extra word.

Example 5.11. For the empirical corresponding to:

$$\mathcal{M}_3 = \{The\ employees\ understood, The\ employees\ understood\ the\}$$

defined above in (5.11) and (5.12), we obtain a signalling fraction of $SF_3 = 0.05$, hence showing that the word *the* at the end of the fragment *The employees understood the* is not difficult to parse. On the other hand, if we calculate the signalling fraction for the empirical model:

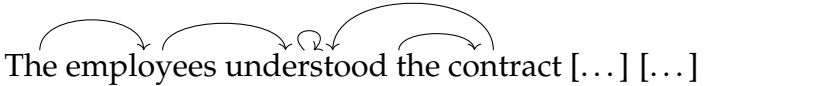
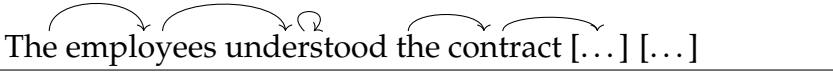
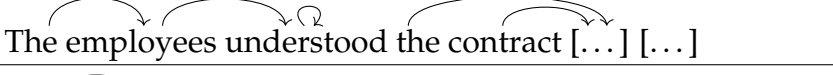
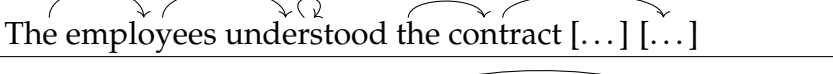
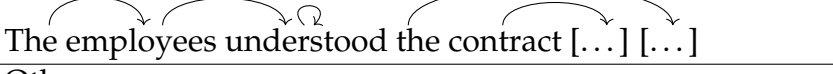
$$\mathcal{M}_5 = \{The\ employees\ understood\ the\ contract, \\ The\ employees\ understood\ the\ contract\ would\}$$

(see Fig. 5.3), the signalling fraction can be found to be $SF_5 = 0.79$, which reflects the fact the parsing the word *would* is quite difficult.


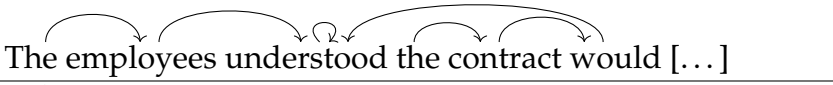
Summary of the chapter

We consider a **sequence of empirical models** such that:

- Each empirical model contains a **pair of contexts** differing by a single word;
- For each context, we select a **probability distribution over possible parses** corresponding to the mental representation of the syntactic structure of a sentence fragment. The details of the computation of the probability distribution will be discussed in the next Chapter.

Parse	Probability
	0.44
	0.14
	0.12
	0.10
	0.04
Other parses	< 0.01

(a) Probability distribution for the context *The employees understood the contract*

Parse	Probability
	0.96
	0.02
Other parses	< 0.01

(b) Probability distribution for the context *The employees understood the contract would*

Figure 5.3: Example of an empirical model corresponding to \mathcal{M}_5 in the sentence *The employees understood the contract would change* (adapted from the empirical model obtained for *The faithful employees understood the technical contract would be changed*, which can be found in [192]).

Although we cannot observe contextuality in these empirical models, we can use the signalling fraction SF to quantify the difficulty of parsing incoming information.

Chapter 6

PREDICTING GARDEN-PATH EFFECTS

In the previous chapter, we introduce our model of the human parsing process. This chapter aims to test the predictions arising from the model using empirical data.

We start by describing the procedure for computing the reading time predictions in Section 6.1. In Section 6.2, we describe the predictions from empirical models, and in Section 6.3, we compare these results with the ones obtained from surprisal theory.

6.1 Methods

In this section, we describe the procedure employed to predict reading times. We start by describing the computational tools used. Subsequently, we explain how we used these tools to collect probabilities. Finally, we describe the datasets from which the garden-path sentences and reading times are taken.

6.1.1 Tools

In this work, we approximate this probability distribution using the large language model BERT and the state-of-the-art dependency parser spaCy.

BERT [46] is one of the first language models to adopt the *transformer* architecture. The transformer was first introduced in [187] in 2017 and offered an alternative to Recurrent Neural Networks (RNNs). This architecture improved both the trainability of neural networks and their performance and is still considered state-of-the-art. See Section 2.1.1 for a more detailed description of BERT.

spaCy is an open-source Python library developed by the company Explosion. It is widely used in NLP, in particular for linguistic annotations. Its functionalities include tokenization, lemmatization, part-of-speech tagging, sentence boundary detection, and dependency parsing. In this work, we mostly made use of the latter. The dependency parser has been evaluated independently in [38] over the OntoNotes5 corpus containing 2.9M tokens. It was shown that the spaCy dependency parser predicted the head of a word (Unlabelled Attachment Score) with a 89.61% accuracy and the head and label of a word (Labelled Attachment Score) with an accuracy of 87.92% [38].

We also worked with different variations of BERT and spaCy to see how the accuracies of the predictions would vary.

For BERT, we used the following flavours:

- `distilBERT`: a light version of BERT. It only has 40% of the parameters of the original `bert-base` model, but runs 60% faster while preserving 95% of its performance accuracies in language understanding tasks;
- `bert-base-cased`: the most commonly used version of BERT. It has 110 million parameters and was trained on the Toronto Book Corpus and the English Wikipedia, both of which distinguish between lower and upper case letters. Uncased versions of the same algorithm exist and were developed for purposes of cross-lingual learning;
- `bert-large-cased`: a larger version of `bert-base-case` which has 340 million parameters.

For spaCy, we also worked with different models, namely:

- `en_core_web_sm`: its standard version. It was trained using convolutional neural networks on web text consisting of blogs, news, and comments;
- `en_core_web_lg`: its larger version. Its training procedure is similar to the `en_core_web_sm` version, but also contains a word vector table with 500k unique 300-dimensional vectors;
- `en_core_web_trf`: its newer version. It has no word vectors but is trained using state-of-the-art transformer-based neural networks.

6.1.2 Method

To obtain a probability distribution over parses, we implement the following procedure:

1. Given a fragment of a sentence S , we turn it into a complete sentence by *masking* all of the remaining words of the S , see Fig. 6.1 for an example.

Remark 6.1. Since the task we are interested in is closely related to the task BERT was trained on, we did not need to fine-tune the pre-trained BERT models.

2. BERT then provides a list of predictions of the completion of the subphrases and a *logit* score σ for each of these predictions, which is meant to rate the likelihood of each prediction. The common practice in NLP is to use the logistic function to turn these scores into probabilities, namely:

$$p = \frac{e^{\sigma}}{1 + e^{\sigma}} \quad (6.1)$$

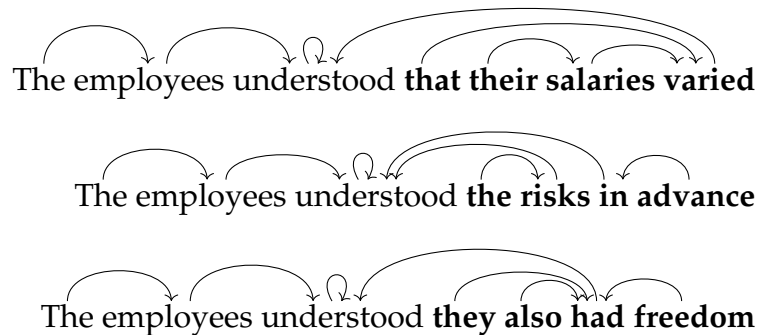
Remark 6.2. We could have equally used a softmax function to convert those scores into probabilities. We leave the investigation of results using softmax as future work.

Now, these predictions are not always words and may include punctuations. We only include the text predictions to avoid complications by dropping the punctuation and renormalising the probability distribution.

Context	BERT input
<i>The</i>	The [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
<i>The employees</i>	The employees [MASK] [MASK] [MASK] [MASK] [MASK]
<i>The employees understood</i>	The employees understood [MASK] [MASK] [MASK] [MASK]
<i>The employees understood the</i>	The employees understood the [MASK] [MASK] [MASK]
<i>The employees understood the contract</i>	The employees understood the contract [MASK] [MASK]
<i>The employees understood the contract would</i>	The employees understood the contract would [MASK]

Figure 6.1: BERT inputs for the sentence *The employees understood the contract would change*.

3. We then use `spaCy` to parse each of the predictions provided by BERT. To obtain the syntactic structure of the specific subphrase we are working with, we restrict the full parse to the words included in that subphrase using the restriction maps from the presheaf \mathcal{E} (see eq. (5.7)).
4. The probability of each such partial parse is obtained by summing up all the BERT-prediction probabilities that restrict to the same parse. For example, the predictions for the continuations of *The employees understood*:



will lead to the same partial parse when restricted to the context *The employees understood*, namely:



Their probabilities will, therefore, be summed up in the corresponding empirical model.

6.1.3 Description of the datasets

In this work, we make use of two reading time datasets that have been collected in psychology, namely the one of Sturt et al. [173] and the one of Grodner et al. [83]. In both of the studies presented in [173] and [83], the authors investigated the slow-downs of garden-path sentences of types NP/S and NP/Z, and both collected (non-cumulative) *self-paced reading* times.

Self-paced reading In self-paced reading experiments, the participants are presented with a sentence or text, where most words are hidden, apart from a text window (usually consisting of a word or several words). Then, upon interaction with a computer (e.g. pressing the space bar), the window is allowed to move from left to right, revealing more text to the participant. In a non-cumulative setting, the

	Display
Step 1	The faithful employees_____
Step 2	_____understood the technical contract_____
Step 3	_____would be changed_____
Step 3	_____would be changed_____
Step 4	_____very soon_____
Step 4	_____very soon_____

Figure 6.2: Evolution of the display presented to participants in a (region-by-region) self-paced reading experiment, with input sentence *The faithful employees understood the technical contract would be changed very soon*

participant cannot access previous text once the window has moved.

Self-paced reading experiments provide less information than eye-tracking experiments do. For example, backtracking is not an option for participants, and they are arguably less natural than the eye-tracking setting. However, they are much more interpretable since they produce fewer variables to keep track of. In addition, they are less expensive to set up since all that is required is a computer, and online crowdsourcing is also possible.

The Sturt et al. dataset

The Sturt et al. [173] dataset consists of 32 pairs of sentences such as:

- (1a) The faithful employees understood the technical contract would be changed very soon.

(1b) Because the faithful employees negotiated the technical contract would be changed very soon.

Each of these pairs contains an NP/S sentence, such as (1a), and an NP/Z sentence, such as (1b), and both the sentences in a pair share overlap in vocabulary.

In addition, each of these garden-path sentences is also associated with *unambiguous version*, which is easier to parse. For NP/S sentences, this is done by adding the connective *that* after the main verb. In NP/Z sentences, this is achieved by adding a comma after the main verb. For example, the following are the unambiguous versions of the sentences (1a) and (1b) respectively:

(1c) The faithful employees understood **that** the technical contract would be changed very soon.

(1d) Because the faithful employees negotiated, the technical contract would be changed very soon.

Remark 6.3. In the following discussion, we sometimes use the term “ambiguous sentences” to describe the garden-path sentences. This is an *abus de langage* since these sentences are not actually (globally) ambiguous. Similarly, the “unambiguous sentences” are not fully (locally) unambiguous. This terminology helps distinguish the sentences that cause difficulty parsing and those that don’t.

This gives a total of 128 sentences.

Each of these sentences is, in turn, divided into 4 regions. For instance, the sentences (1a) and (1b) are respectively divided as:

(1e) The faithful employees / understood the technical contract / would be changed / very soon.

(1f) Because the faithful employees / negotiated the technical contract / would be changed / very soon.

and similarly for the unambiguous sentences. The critical regions are the ones underlined.

The experiment described in [173] recorded the *region-by-region* reading times, i.e. the participants were presented with one of these regions at a time and allowed to move one region forward at each event. The numbers reported in the study were

average region reading times, averaged across sentences of the same type (i.e. NP/S ambiguous, NP/S unambiguous, NP/Z ambiguous, or NP/Z unambiguous), and across participants. These numbers can also be found in Table 6.1.

	Regions			
	1	2	3	4
NP/S (ambiguous)	990	1183	877	771
NP/S (unambiguous)	981	1282	790	768
NP/Z (ambiguous)	914	1269	1335	848
NP/Z (unambiguous)	998	1384	935	832

Table 6.1: Region-by-region self-paced reading times of garden-path sentences and their unambiguous variants (in ms). These numbers were taken from [173]

The Grodner et al. dataset

The Grodner et al. dataset [83] was created in a different way. As for the Sturt et al. dataset, it contains both NP/S and NP/Z sentences such as :

- (2a) The employees understood the contract would be changed to accommodate all parties.
- (2b) Even though the band left the party went on for at least another two hours.

In addition, each of the NP/S and NP/Z sentences also come with a *modified variant*, where a descriptive noun-phrase is added to the subject of the main verb. For example:

- (2c) The employees **who initiated the strike** understood the contract would be changed to accommodate all parties.
- (2d) Even though the band **which played funk music** left the party went on for at least another two hours.

Remark 6.4. The reason for having modified and unmodified versions of the same garden-path sentence was to identify whether human parsing strategies were more consistent with repair-based or reanalysis-based models (see Section 2.2.1). However, since we are not interested in this particular aspect of the parsing process, we ignored their distinction when analysing the garden-path effects. However, when

creating a linear regression model, having a multiplicity of data points allows us to obtain a better model and reduces errors due to averaging.

The dataset taken from [83] contains 20 pairs of modified/unmodified NP/S garden-path sentences and 20 pairs of modified/unmodified NP/Z sentences.

In addition, as for the Sturt et al. dataset, each sentence comes with an unambiguous version. For instance, for the unmodified versions (2a) and (2b), these are:

(2a) The employees understood **that** the contract would be changed to accommodate all parties.

(2b) Even though the band left the party, went on for at least another two hours.

Similarly, for the modified version (2c) and (2d), the unambiguous variants are:

(2a) The employees who initiated the strike understood **that** the contract would be changed to accommodate all parties.

(2b) Even though the band which played funk music left the party, went on for at least another two hours.

This gives us a total of 160 sentences.

Each of these sentences is also divided into regions, but contrary to the dataset of [173], the number of regions is different for every type of sentence, and the length of regions is highly variable within a given sentence. These regions are depicted in Table 6.2a.

In the corresponding study, the authors collected *word-by-word* self-paced reading times and reported the average word reading time for each region (aside from the last one for which numbers are omitted), averaged across sentences of the same type and across participants. The obtained averages are shown in Table 6.2b.

Remark 6.5. In addition, since the numbers quoted in the Sturt et al. dataset are region-by-region reading times, we decided to make region-by-region predictions over this dataset. Similarly, since the Grodner et al. dataset used word-by-word reading times, we decided to make word-by-word predictions over this dataset. This is at odds with the study of [185] where, for uniformity purposes, the region-by-region reading times were averaged to produce word-by-word reading times. However, doing so would increase the amount of systematic error. For instance,

	Regions					
	1	2	3	4	5	6
NP/S (unmod., amb.)	The employees	understood	the contract	would be changed		
NP/S (unmod., unamb.)	The employees	understood	that	<u>the contract</u>	would be changed	
NP/S (mod., amb.)	The employees	who initiated the strike	understood	the contract	<u>would be changed</u>	
NP/S (mod., unamb.)	The employees	who initiated the strike	understood	that	the contract	would be changed
NP/Z (unmod., amb.)	Even though the band	left	the party	<u>went on for[...]</u>		
NP/Z (unmod., unamb.)	Even though the band	left,	the party	<u>went on for[...]</u>		
NP/Z (mod., amb.)	Even though the band	which played funk music	left	the party	<u>went on for[...]</u>	
NP/Z (mod., unamb.)	Even though the band	which played funk music	left,	the party	<u>went on for[...]</u>	

(a) Regions of the different sentence types in the Grodner et al. dataset. The critical regions are underlined. (Note that the last region is omitted)

	Regions					
	1	2	3	4	5	6
NP/S (unmod., amb.)	397	467	412	424		
NP/S (unmod., unamb.)	393	460	431	396	410	
NP/S (mod., amb.)	392	415	449	401	419	
NP/S (mod., unamb.)	398	413	471	393	388	391
NP/Z (unmod., amb.)	452	402	382	452		
NP/Z (unmod., unamb.)	400	452	402	383		
NP/Z (mod., amb.)	433	407	464	415	432	
NP/Z (mod., unamb.)	405	400	494	448	395	

(b) Average word-by-word self-paced reading times of garden-path sentences and their unambiguous variants (in ms). (numbers taken from [83])

Table 6.2: Description of the Grodner et al. dataset

assuming that our signalling fraction SF is indeed related to reading times and assuming that the buffering time associated with the change of stimulus of the screen is constant, this extra time (unrelated to the reading difficulty) is only added once per region in the region-by-region setting, but multiple times per region in the word-by-word setting. Hence, this buffering time is constant in the former but dependent on the length region in the latter. Due to these differences in the reading time collection process in the two datasets, we decided to present the predictions obtained for the Sturt et al. dataset and the ones for the Grodner et al. dataset separately.

6.2 Analysis of the predictions

In this section, we investigate the prediction power of the signalling fraction SF. We start by looking at the empirical correlation between SF and reading times. Then, we create linear regression models of reading times from the signalling fraction. Using this model to produce prediction, we then investigate whether we can observe

a garden-path effect and whether we can see a difference in this garden-path effect between NP/S and NP/Z sentences. In the following section, we will compare our results with the existing ones in the literature that use surprisal.

Remark 6.6. For the remainder of this chapter, we will study the *linear correlations* between SF and reading times. We made this choice for simplicity and not according to any heuristic. Investigating other types of relations between these quantities is left to future work.

6.2.1 The Sturt et al. dataset

The linear regression model

Starting from the assumption that the signalling fraction SF of an empirical model with contexts $\mathcal{M} = \{m, mw\}$ correlates with the amount of difficulty induced by reading the word w (see Section 5.2). From this, we expect that region reading time correlates with the sum of the signalling fractions, summed over all the words in the region. Now, given that we are studying linear relations between the signalling fractions and reading times, we expect this relation to be of the form:

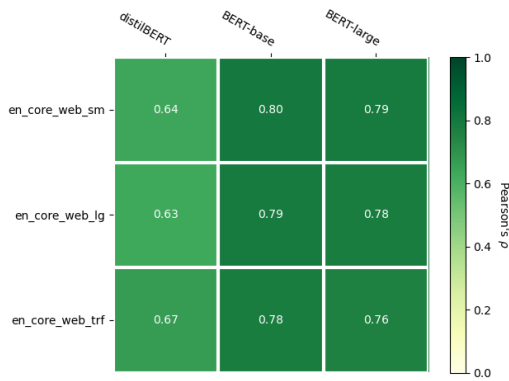
$$RT(r) = \alpha \sum_{w \in r} SF(w) + \beta \quad (6.2)$$

for any region r and w in that region. In the above equation, we denote by $SF(w)$, the signalling fraction associated with the empirical model with contexts $\mathcal{M} = \{m, mw\}$.

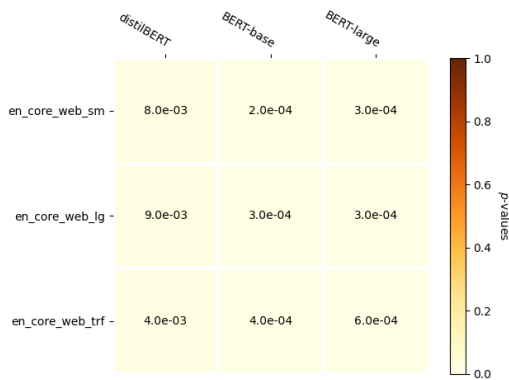
Moreover, since [173] presented the average reading time over sentences of the same type, we thus can check whether the following holds:

$$\langle RT(r(\mathbf{S})) \rangle_{\mathbf{S}} = \alpha \left\langle \sum_{w \in r(\mathbf{S})} SF(w) \right\rangle_{\mathbf{S}} + \beta \quad (6.3)$$

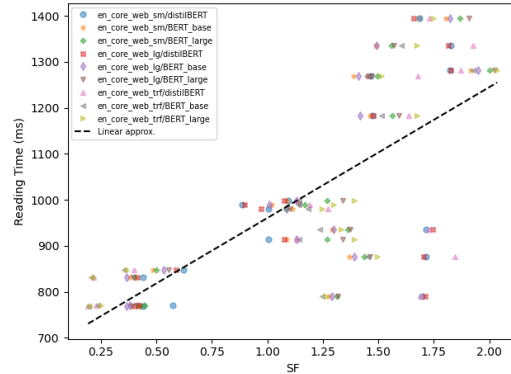
where this time, the region $r(\mathbf{S})$ denotes a particular region of a given sentence \mathbf{S} .



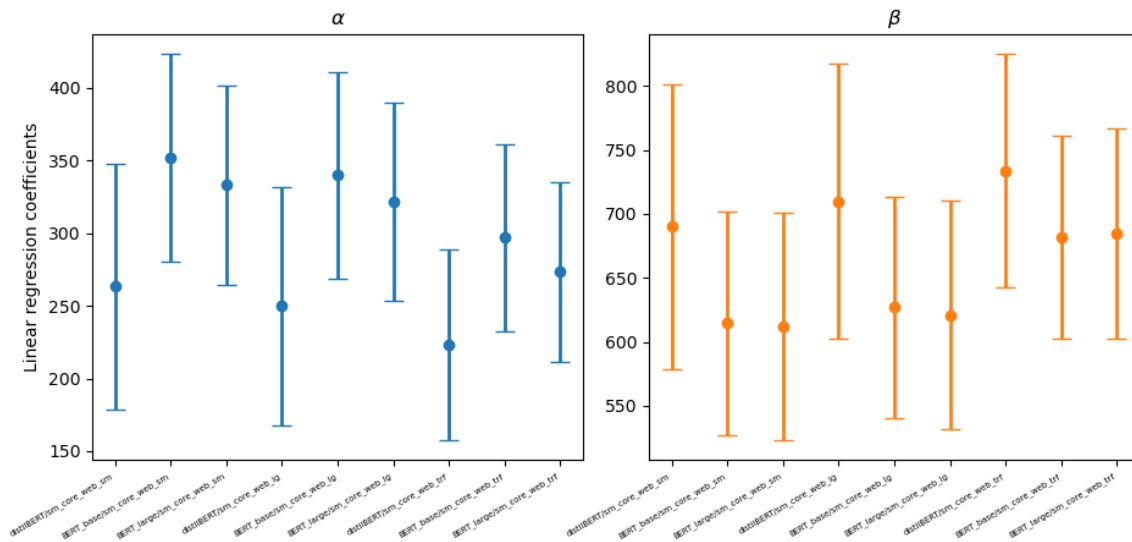
(a) Pearson's ρ coefficients.



(b) p -values associated with the Pearson's ρ coefficients.



(c) Linear correlation between SF and self-paced reading times.



(d) Coefficients of the linear regressions obtained for different BERT and spaCy variants. The standard error on these coefficients is depicted as error bars.

Figure 6.3: Analysis of the linear correlations between SF and reading times in the Sturt et al. dataset.

Correlation with reading times

To test the above hypotheses, we calculate Pearson’s ρ coefficients associated with SF and reading times. The obtained ρ coefficients and their associated p -values (testing the confidence level at which we have $\rho \neq 0$) are shown in Fig. 6.3a and 6.3b respectively.

As we can see, these correlation coefficients are generally high (all of them > 0.63) and, importantly, all positive. Furthermore, the p -values associated with the correlation coefficients are statistically significant ($p < 9 \times 10^{-3}$ for any choice of BERT and spaCy variants). These results are evidence of a robust monotonic relation between SF and reading times. I.e. when SF increases, so does the reading time.

Impact of the different BERT and spaCy variants In addition, these coefficients are higher ($\rho \simeq 0.78$) for any empirical model obtained from the larger BERT-base or BERT-large as compared to the empirical models obtained from the lighter version distilBERT. Similarly, the p -values appear to be larger for models using distilBERT (of the order of magnitude of $p \sim 10^{-3}$) as opposed to ones obtained from the other BERT variants ($p \sim 10^{-4}$). This suggests that reducing the parameters of BERT may impact our performance accuracies.

There is, however, no sign of the influence of the spaCy models by solely looking at the ρ coefficients.

Regression models From the existence of a linear correlation, the use of the linear model of (6.3) is justified. The coefficients α and β calculated for empirical models calculated from different BERT and spaCy variants are shown in Fig. 6.3d. We can then see that the obtained coefficients are comparable for all of the BERT and spaCy models, which overall give a linear regression model around:

$$\langle RT(r(\mathcal{S})) \rangle_{\mathcal{S}} \simeq 295 \left\langle \sum_{w \in r(\mathcal{S})} \text{SF}(w) \right\rangle_{\mathcal{S}} + 664 \quad (6.4)$$

For the rest of this work, however, we will take the individual regression models obtained for each of the BERT and spaCy variants. This will then ensure that we get the best possible predictions.

Predicting garden-path effects

Using these regression models, we can make predictions of reading times from SF. In turn, we can evaluate these predictions' accuracy by looking at what effect they can and cannot predict.

Methods We start by investigating whether SF can predict garden-path effects, i.e. whether the reading times of garden-path sentences are higher than the reading times for the equivalent unambiguous sentences (over their critical region).

To do so, we calculate the so-called *garden-path effect* of a garden-path sentence by simply taking the difference in reading time of the critical region in the ambiguous and unambiguous versions, i.e.:

$$GPE(S) = RT(r_{critical}(S)) - RT(r_{critical}(unambiguous(S))) \quad (6.5)$$

where $r_{critical}(S)$ isolates the critical region of S , and $unambiguous(S)$ gives the unambiguous version of a sentence S .

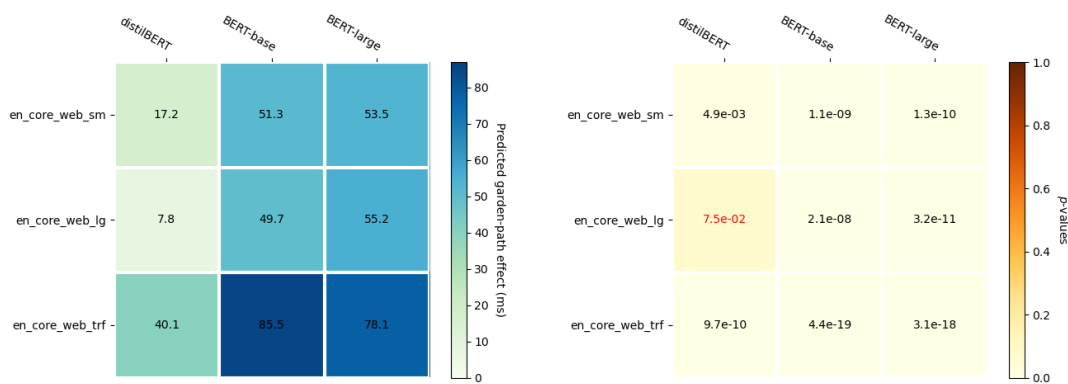
Results The average garden-path effect obtained for empirical models using the different variants of BERT and spaCy are shown in Fig. 6.4a. We can observe that, on average, these predicted garden-path effects are indeed positive, therefore showing that SF predicts higher reading times for garden-path sentences than their unambiguous versions.

To strengthen this result, we also conducted 1-sample t -tests testing the null hypothesis that this average is 0 (i.e. that the reading time predictions are the same for ambiguous and unambiguous sentences). The resulting p -values are depicted in Fig. 6.4b. We can see all of the p -values are indeed statistically significant, except for the empirical models using distilBERT and the en_core_web_lg pipeline of spaCy (where even the p -value is relatively low and is $p = 0.07$).

These results overall show that SF can confidently detect the existence of a garden-path effect.

NP/S and NP/Z predictions

We now want to analyse the prediction for garden-path effects for NP/S and NP/Z sentences separately.



(a) Average predicted garden-path effect. (b) p -values associated with the 1-sample t -tests evaluating whether the average garden-path effect is 0.

Figure 6.4: Analysis of the predicted garden-path effect over the Sturt et al. dataset.

The distributions of the obtained garden-path effects for NP/S and NP/Z sentences are shown in Fig. 6.5. As we can see, SF overall underestimates the garden-path effects, particularly for NP/Z sentences. This suggests that SF alone does not entirely explain the full difficulty of garden-paths. However, we can also observe that increasing the number of parameters of both the BERT and `spaCy` models improves the predictions.

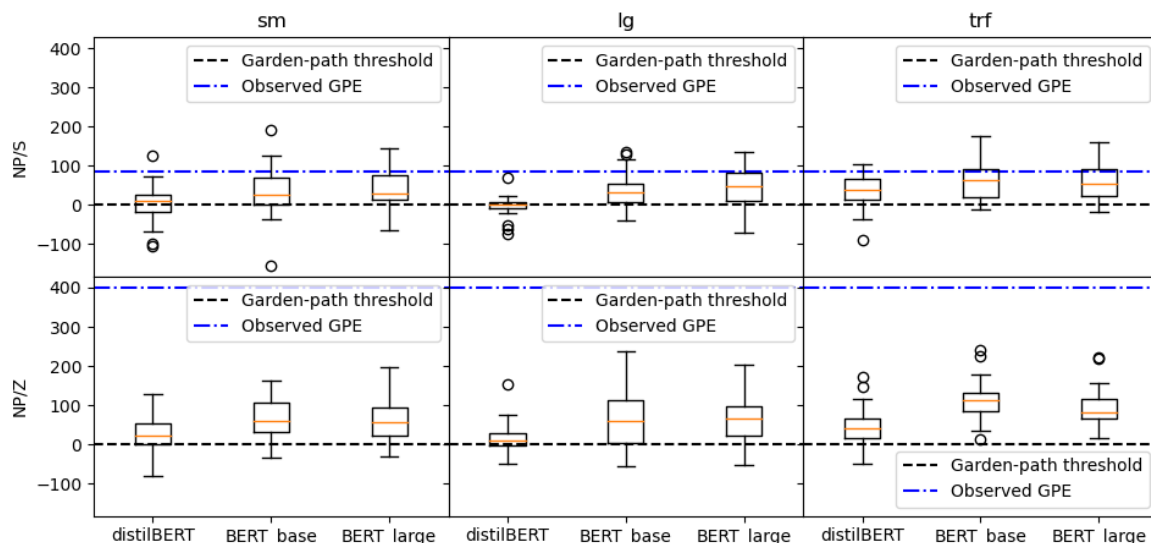


Figure 6.5: Boxplots of the garden-path effects predicted over the Sturt et al. dataset. The human baseline is also quoted.

Finally, we are also interested to see whether SF can detect different levels of difficulty, namely between the NP/S and NP/Z sentences. By comparing the average predicted garden-path effects of NP/S and NP/Z sentences shown in Figs. 6.6a and 6.6b, one can see that the garden-path effects are generally higher for NP/S than for NP/Z sentences.

We then tested this hypothesis by conducting *t*-test comparing them and found that this difference is statistically significant for most BERT and spaCy variants. Surprisingly, the empirical models using BERT-large did not perform well under these *t*-tests. This negative result could still be due to the noisiness of human data or the several averaging steps that have occurred to obtain data in the first place.

On the whole, this gives us evidence that SF can identify different levels of parsing difficulty.

6.2.2 The Grodner et al. dataset

The linear regression models

We now analyse the Grodner et al. dataset predictions. We first recall that the figures quoted in [83] are not region-by-region but word-by-word reading times, averaged for each region across different sentences. Hence, instead of finding a linear regression model of (6.3), we will be interested in models of the form:

$$\langle RT(w) \rangle_{w \in r(s), s} = \alpha \langle SF(w) \rangle_{w \in r(s), s} + \beta \quad (6.6)$$

Correlation with reading times

As we did for the Sturt et al. dataset, we test this hypothesis by first computing the associated Pearson's ρ coefficients and associated *p*-values. These can be found in Fig. 6.7a and 6.7b respectively.

The obtained correlation coefficients are found to be smaller than the ones obtained for the Sturt et al. dataset (here, $\rho > 0.35$ for all BERT and spaCy variant). The correlation is still positive, and we achieve correlations up to $\rho = 0.56$. Similarly, the *p*-values are generally higher than the ones of Fig. 6.3b, but still statistically significant (all of the $p < 0.04$).

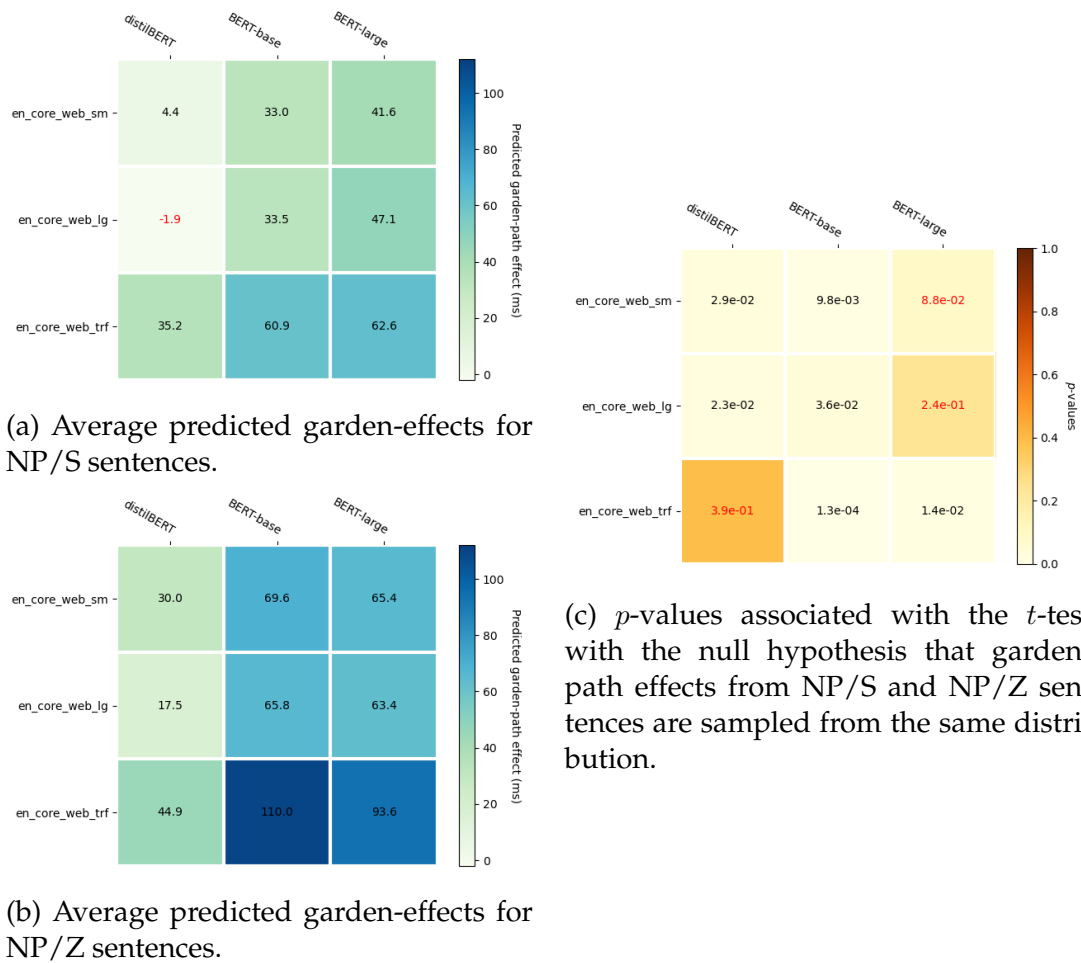
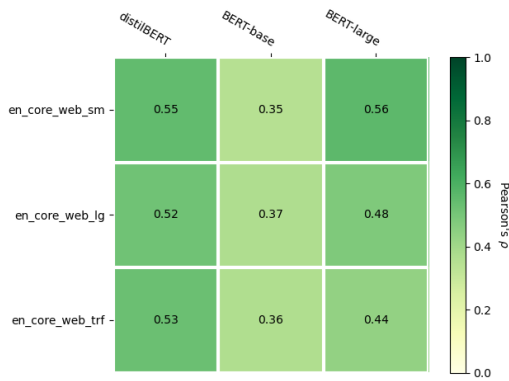


Figure 6.6: Comparison of the predicted garden-path effects between NP/S and NP/Z sentences.

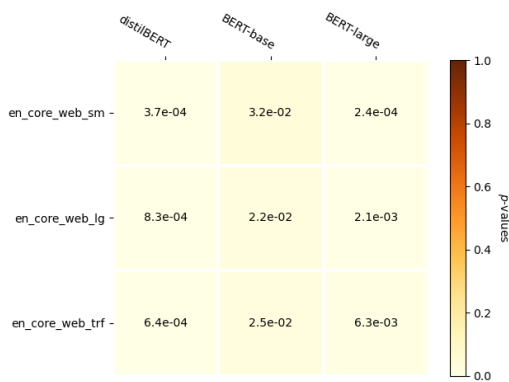
Impact of the BERT and spaCy variants Similarly to the other dataset, we observe that the coefficients ρ seem to be more affected by the choice of BERT model than the choice of spaCy variant. However, contrary to the previous results, we observe that it is the BERT-base model that led to the worse correlations and that the distilBERT empirical models lead to fairly high correlations ($\rho \sim 0.53$).

Overall, the positive correlations and the low p -values strengthen our previous findings that SF are correlated with reading times.

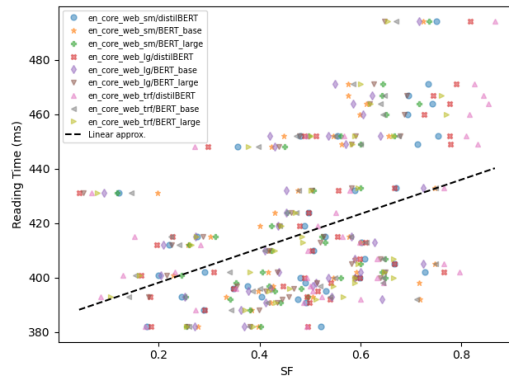
Linear regression As before, we can use linear regression equations to predict reading times from signalling fractions. The different α and β coefficients obtained for different choices of BERT and spaCy variants are shown in Fig. 6.7d.



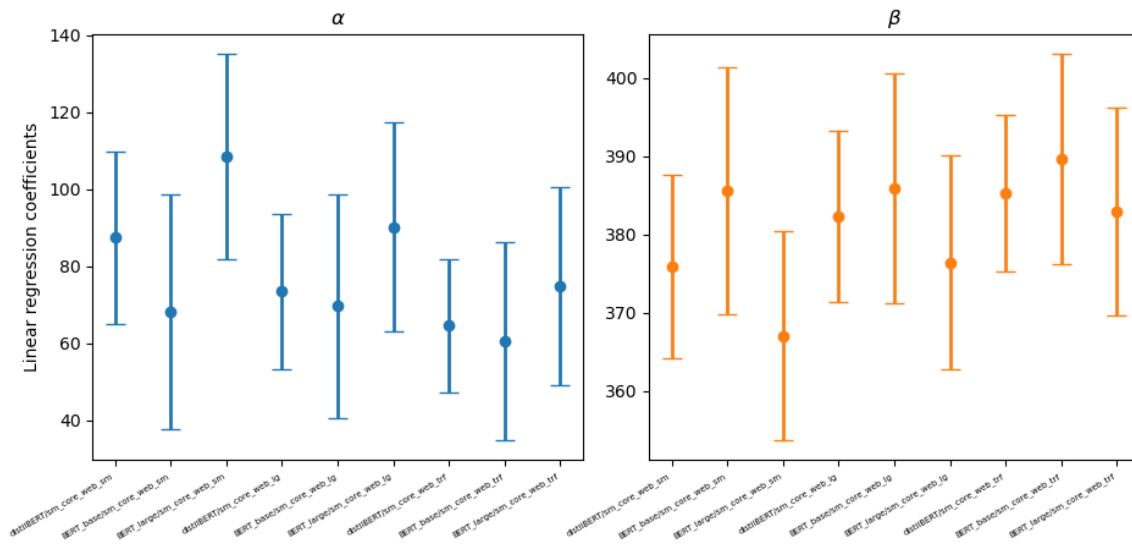
(a) Pearson's ρ coefficients.



(b) p -values associated with the Pearson's ρ coefficients.



(c) Linear correlation between SF and self-paced reading times.



(d) Coefficients of the linear regressions obtained for different BERT and spaCy variants. The standard error on these coefficients is depicted as error bars.

Figure 6.7: Analysis of the linear correlations between SF and reading times in the Grodner et al. dataset.

These coefficients are also highly similar and revolve around the following model:

$$\langle RT(w) \rangle_{w \in r(s), s} = 77 \langle SF(w) \rangle_{w \in r(s), s} + 381 \quad (6.7)$$

As done previously, we use each linear regression model for the rest of the analysis to reduce the number of errors due to averaging.

Predicting garden-path effects

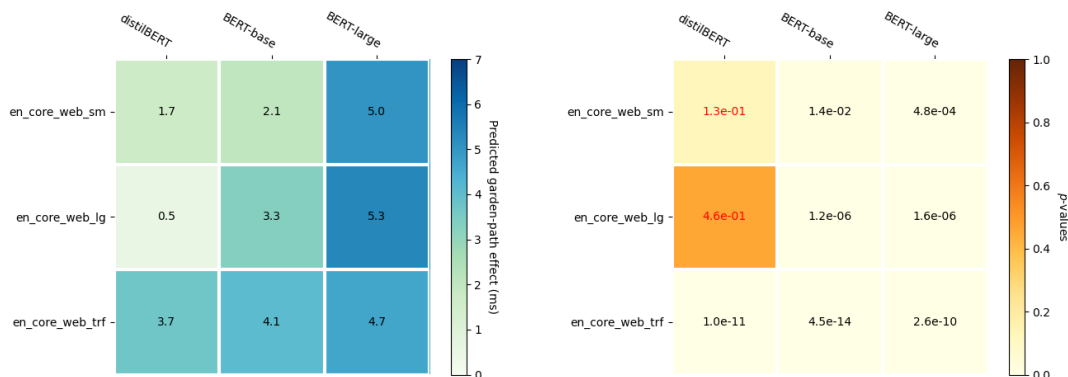
We then compute the predictions of the garden-path effect. Similarly to (6.5), we calculate this garden-path effect as:

$$GPE(s) = \langle RT(w) \rangle_{w \in r_{critical}(s)} - \langle RT(w) \rangle_{w \in r_{critical}(unambiguous(s))} \quad (6.8)$$

Results As before, we want this garden-path effect to be greater than 0. This is indeed the case (see Fig. 6.8a).

We furthermore conducted 1-sample t -tests to quantify the confidence of this finding and found that these garden-path effects are indeed statistically significantly non-zero, aside from the empirical models using `distilBERT` along with the `spaCy` models `en_core_web_sm` and `en_core_web_lg`.

We therefore conclude that SF can also detect a garden-path effect in this dataset.



(a) Average predicted garden-path effect. (b) p -values associated with the 1-sample t -tests evaluating whether the average garden-path effect is 0.

Figure 6.8: Analysis of the predicted garden-path effect over the Sturt et al. dataset.

NP/S and NP/Z sentences

Now focusing on the difference in predictions between NP/S and NP/Z sentences, we depict the distributions of predicted garden-path effect for the two categories of garden-path sentences in Fig. 6.9.

We observe that, as before, SF underestimates the garden-path effect of both NP/S and NP/Z sentences. In addition, in some NP/S empirical models, the predicted garden-path effect is *negative*. The models showing a negative garden-path effect are, however, restricted to the small BERT (distilBERT or BERT-base), and spaCy (en_core_web_sm or en_core_web_lg) models. In the larger models, the predicted garden-path effects become positive.

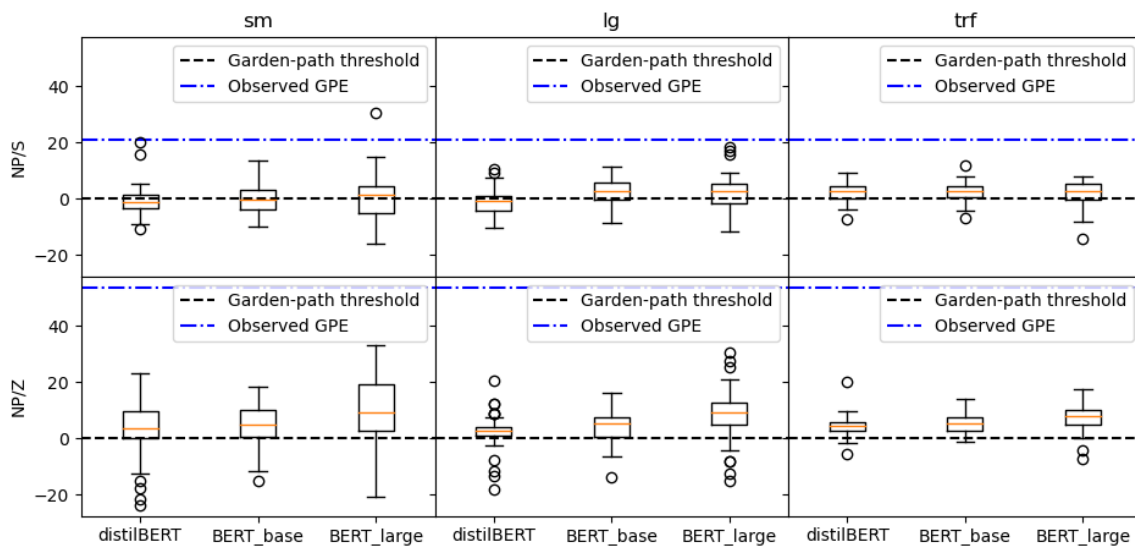


Figure 6.9: Boxplots of the garden-path effects predicted over the Sturt et al. dataset. The human baseline is also quoted.

It also appears that, as in the Sturt et al. dataset, the average garden-path effects are higher for NP/Z sentences than for NP/S sentences. This effect is shown empirically (see Figs. 6.10a and 6.10b).

Moreover, we can estimate the confidence levels of the claim that SF distinguishes between NP/S and NP/Z sentences using t -tests. The obtained p -values are depicted in Fig. 6.10c. These t -test are all statistically significant aside from the em-

pirical models using BERT-base in conjunction with `en_core_web_lg` pipeline of `spaCy` (and even then, the p -value is found to be $p < 0.1$).

Contrary to the ongoing trend, these p -values are more statistically significant than the ones obtained for the Sturt et al. dataset, therefore showing that the difference between garden-path effect predictions of NP/S and NP/Z sentences is *more marked* in the Grodner et al. dataset.

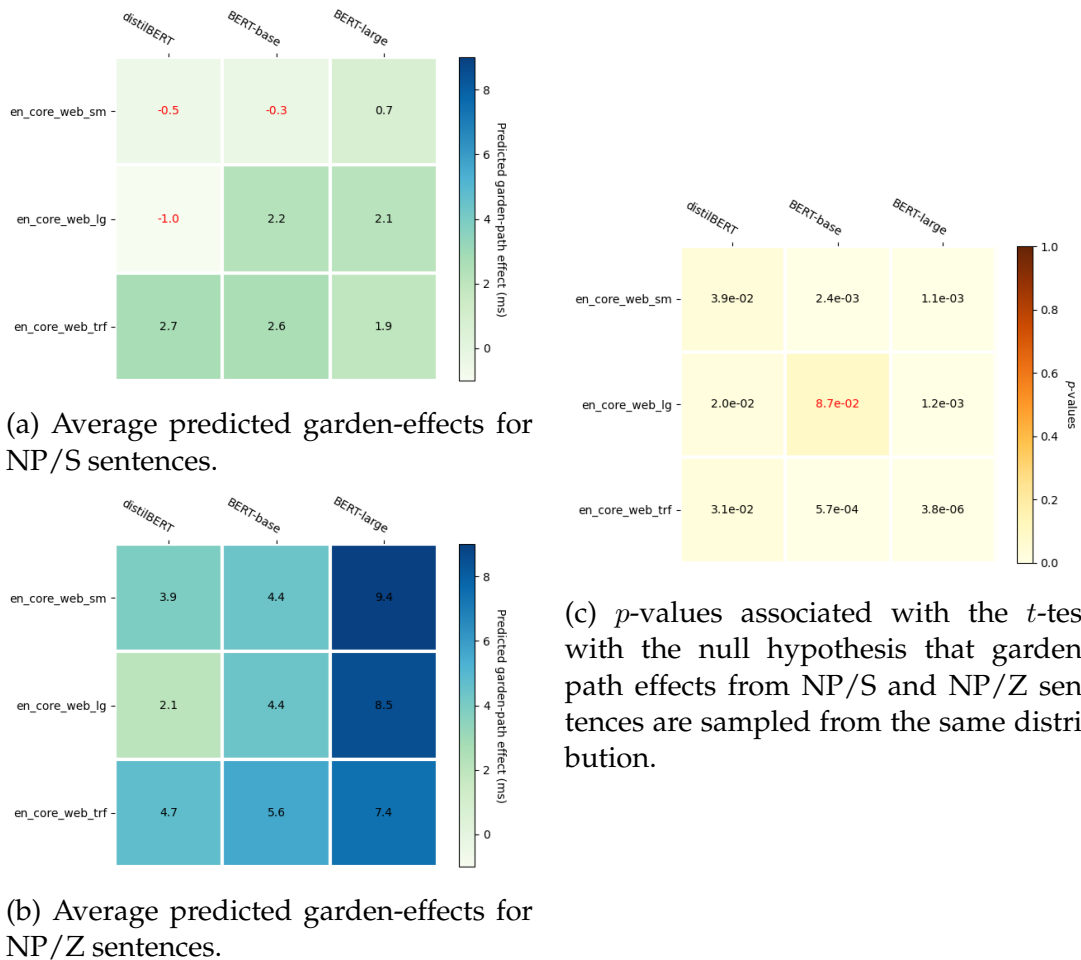


Figure 6.10: Comparison of the predicted garden-path effects between NP/S and NP/Z sentences.

6.2.3 General discussion

By quantifying the amount of signalling in certain empirical models, we have obtained an alternative measure of the difficulty of parsing. We have seen that the

signalling fraction SF does have a strong positive correlation with human reading times. Using a linear regression model lets us obtain predictions for reading times and garden-path effects. We have also found that this predicted garden-path effect is (statistically significantly) higher for NP/Z sentences than for NP/S sentences. Finding such differences has yet to be shown to be possible to identify using surprisal theory only [185, 186, 95].

Although SF correctly identifies the relative difficulty levels of different garden-path sentences, the overall magnitude of the predicted effect is systematically underestimated. This is similar to the findings from surprisal theory [185, 186, 95, 14], and we therefore have similar hypotheses to explain why this is the case.

For instance, this could be evidence that the human parsing process is purely forward-looking. Many psycholinguistic theories suggest that backtracking or re-processing is necessary to make sense of garden-path sentences, both of which are backward-looking processes. However, there is no clear evidence that this is indeed the case. The field of research is globally more interested in the cause of difficulty (e.g. vocabulary biases, plausibility, ...) and the nature of the parsing strategies (e.g. presence or absence of backtracking, parallel or sequential parsing, ...), not the algorithmic parsing procedure (i.e. the specific steps the reader carry out when parsing a sentence), which is not even guaranteed to be the same for all individuals.

In addition, part of the difficulty also comes from the nature of the datasets used in this experiment. Indeed, only a few averaged data points were available to compute the linear regression models. Hence, taking those averages would have considerably increased the amount of error even before any computation was done. In particular, we have observed that the predictions obtained from the Grodner et al. dataset were almost consistently worse than those obtained from the Sturt et al. dataset. From these considerations, this could be because one more round of averaging was done in this dataset (i.e. averaging of the reading times for each region).

Other explanations for this discrepancy could be due to the unevenness of the dataset. For instance, the number of characters per region, the number of words per region, and the vocabulary were not specifically controlled. To remedy this, the authors of [83] decided to use a linear regression model tailored to each participant to normalise the reading times and determine whether a word is read slower or faster than expected. These normalised reading times were referred to as *residual*

reading times. However, the parameters of these linear regressions are not available, and it was still more convenient to use raw reading times in our study rather than residual reading times.

To deal with those issues, we plan to use more detailed datasets in future works, such as self-paced reading time datasets collected in [151, 95], which contain word-by-word reading times. Using a more detailed dataset also comes with its drawbacks. For instance, *spillover* (i.e. delay in the observed difficulty) has to be accounted for. However, workarounds exist in the literature [186, 95]. Our first steps will then use them.

In addition, although surprisal is not by default capable of studying backward-looking processes, only a few modifications to the procedure described in Section 5.1 are necessary to be able to account for more complex parsing strategies. We expand this in more detail in the conclusion.

6.3 Comparison with surprisal

We now want to put our results in perspective and compare them with the state-of-the-art methods from computational linguistics, which use surprisal theory. To simplify this comparison, the work presented in [185] used the same datasets as we did to produce garden-path effect predictions from surprisal.

Remark 6.7 (On the fairness of the comparison). There are several reasons why the comparison between the predictions of [185] and the ones presented in Section 6.2 may not be fair. Firstly, the calculations were not using the same language models. The authors of [185] trained their own LSTM from the Wikipedia (2 million and 90 million tokens versions) and Wall Street Journal corpora. In contrast, we used the transformer model BERT (which did not exist when [185] was published). The LSTMs trained in [185] are not openly available. Therefore, it wasn't easy to obtain a fair comparison, and using pre-trained models was by far the most convenient solution for us.

Magnitude of the garden-path effects

We start by comparing the magnitude of the garden-path effect predictions from surprisal and SF.

The best predictions from Section 6.2, as well as the best predictions from [185] is presented in Table 6.3.

We observe that SF outperforms surprisal considerably better in the Sturt et al. dataset, where the predictions are up to 40% more accurate than the ones of [185] for NP/S sentences and about 20% more accurate for NP/Z sentences.

On the other hand, the surprisal predictions over the Grodner et al. dataset appear to be more accurate than the ones obtained using SF. However, this accuracy decrease is only 20% for NP/S sentences and 3% for NP/Z sentences.

What is quite interesting is that the study of van Schijndel and Linzen in [185] reported that surprisal performed much better over the Grodner et al. dataset. In contrast, our investigation led to more accurate predictions over the Sturt et al. dataset. This discrepancy could explain why our results are better in the Sturt et al. dataset, whereas surprisal outperforms SF over the Grodner et al. dataset.

However, the cause of such discrepancy in accuracies is not clear. This suggests that surprisal and SF do not give the same weights to the same features.

		Prediction (ms)		Observed (ms)
		SF	<i>S</i>	
Sturt et. al	NP/S	62.6	24*	87
	NP/Z	110	30*	400
Grodner et. al	NP/S	2.73	7	21
	NP/Z	8.52	10	53.5

Table 6.3: Comparison of the garden-path effects obtained using surprisal (numbers taken from [185]) and SF. *These numbers have been converted to be a region reading time from the word-by-word reading times quoted in [185].

NP/S and NP/Z predictions

Our more clear-cut results were regarding the difference in garden-path effect predictions for sentences with different levels of difficulty.

Indeed, even though *p*-values were not quoted in the various studies using surprisal for predicting garden-path effects [185, 186, 95, 14], the authors identify the main issue with surprisal as not being able to distinguish between NP/S and NP/Z sentences. In fact, the trend observed in a follow-up study was that NP/S garden-path effects were, on average, higher than the predictions for NP/Z sentences [186].

In conjunction with consistent underestimation of the slowdown prediction, this is their primary motivation for advocating backward-looking mechanisms in parsing strategies.

The fact that we can find such statistical differences using a forward-looking model does not invalidate this hypothesis. After all, our predictions are still widely underestimating the slowdowns as well. However, this may show that there might be some features that surprisal cannot detect, which opens up the question of what other quantities (even apart from SF) could contribute to the reading difficulty of garden-path sentences.

Linking SF and surprisal

Even though our usage of SF stemmed from similar motivations to those for surprisal, it is unclear whether they are mathematically related. The reason for the better performance of SF is that surprisal, as used in this [185], mostly focuses on lexical items. In contrast, syntactic structures are first-class citizens for the SF quantity described here.

Only very recently, the role of syntactic structure in conjunction with surprisal has come into light: in [14], it was shown that syntactic surprisal performs slightly better than pure lexical surprisal but still falls short when distinguishing NP/S from NP/Z and the differences in garden-path effects.

The results of [14] and the ones presented here motivate the hypothesis that syntactic structures are the main deciding factor in the difficulty of garden-path sentences.

Another aspect of our work, which may have led to more accurate results, is that our model can take long-distance dependencies into account, whereas surprisal is not.

Summary of the chapter

This Chapter used our sheaf-theoretic models to predict reading times and garden-path effects. Using two datasets from the psycholinguistic literature, we obtained the following results:

- The correlation between SF and reading times was positive and statistically significant;
- Using a linear regression model, we successfully predicted the existence of garden-path effects. However, we consistently underestimated the magnitude of the effect.
- We accurately predicted that the garden-path effects were higher for NP/Z than for NP/S sentences.

The signalling fraction clearly outperforms the surprisal predictions of the Sturt et al. dataset, both by the magnitude and by distinguishing the NP/S and NP/Z sentences. Over the Grodner et al. dataset, surprisal achieves more accurate predictions than SF, but does not distinguish between NP/S and NP/Z sentences.

CONCLUSION

In this thesis, we studied natural language data from the perspective of foundational quantum mechanics.

We observed how contextuality arises in natural language data and found uses for various quantities in psycholinguistics. To this end, we used the causal and signalling fractions, which were so far only used to describe quantum systems. Our results demonstrated that the sheaf-theoretic framework of contextuality and causality does uncover some linguistic phenomena relating to lexical and syntactic ambiguity arising from human behaviour.

Lexical Ambiguity We started by looking at lexical ambiguities, where the analogy between words and quantum systems appeared more natural. Our detailed analysis showed that although contextuality is hard to obtain in the statistics of lexically ambiguous phrases, it is still possible to find witnesses of quantum-like contextuality, as defined under the Contextuality-by-Default framework. This is evidence of the essential role of the context in the disambiguation process of lexically ambiguous items. In addition, we also saw that the causal fractions of SV and VO empirical models confirmed that the observed probability distributions were primarily consistent with verb after subject and verb after object disambiguation orders. This finding showed that verbs tend to be disambiguated after their arguments, which is consistent with the psycholinguistic theories of the disambiguation of lexically ambiguous words. Using this finding, we simulated the lexical disambiguation process using variational quantum circuits. We also demonstrated that these circuits could, in turn, predict the different interpretation probability distributions associated with unseen phrases. This last result is exciting as it only required a small training set; in theory, only knowledge of 8 probability distributions should be able to predict

the probability distributions of 8 new ones, and each of these probability distributions only required annotations of 25 participants, which is much lower than the resources needed to train a large language model.

To our knowledge, this project is the first to study the disambiguation process of phrases containing two target words of different grammatical roles and with explicit syntactic relations between them. Moreover, the proof-of-principle that variational quantum circuits can simulate the (human) lexical disambiguation process opens several questions regarding their performance in standard NLP tasks, including word-sense disambiguation. In addition, although we have only focused on two possible interpretations of each word, the approach can easily be extended to an arbitrary number of possible interpretations.

Syntactic ambiguity We then turned our attention to syntactic ambiguities and notably focused on particular sentences, namely garden-path sentences, that are important for studying the human syntactic parsing process.

We first observed high correlations between the signalling fraction and reading times of garden-path sentences. This result showed that the signalling fraction correlates with the difficulty of parsing a given sentence fragment. We then use such correlation to produce a linear model of reading times in terms of the signalling fraction. This linear regression model allowed us to predict reading times and garden-path effects associated with different sentences. These predictions outperformed the current state-of-the-art predictions of computational linguistics using surprisal theory. Among these, the most crucial improvement from SF was to find statistically significant differences in NP/S and NP/Z sentences, where the former is significantly easier to parse than the latter. This may be evidence that our (significantly simplified) parsing model may be closer to the actual human parsing process than surprisal theory.

We believe this project paves the way for better quantum-based NLP algorithms, as it sheds some light on how different aspects of natural language ambiguities would benefit from quantum advantages. The mathematical frameworks we used led to meaningful results and provided proof of concept that they can be used to talk about linguistic phenomena.

Future work

This approach adopted in this project offers more possible research lines. We here describe a few of the possible extensions of this project.

General improvements

The first and most obvious way to extend our approach is to loosen some simplifications imposed on the different empirical models. For instance, it will be worth expanding our lexical ambiguity empirical models to include all of the possible interpretations of each word, e.g. using its different `WordNet` senses. Similarly, it would be interesting to see whether adding the labels of the dependencies in syntactic empirical models would impact the accuracy of the reading time predictions. Furthermore, we could also consider expanding from having two possible choices of words for subjects, verbs, and objects to having the full vocabulary in lexical ambiguity models. In the case of the syntactic models, we could consider all possible ways a sentence fragment can be completed instead of restricting ourselves to the observed continuation. By doing so, we may have a closer link with surprisal theory.

In addition, it will be interesting to combine the syntax and semantics models, which were described independently in Parts II and III. At the moment, we can envisage two ways of doing so. The first would consist of concatenating syntactic and semantic empirical models and possibly adding an ad-hoc notion of interaction between the two. On the other hand, the psychology literature suggests that syntactic information has more influence on the semantic level than the other way around. Hence, another (more complex) possibility would be to have a higher-order causal order in which syntactic empirical models could influence any semantic process (but not necessarily the other way around). We could extend this further by considering other knowledge sources, such as pragmatic information and plausibility.

Lexical Ambiguity

Proof of quantum advantage Regarding lexical ambiguity empirical models, even though we demonstrated the existence of contextual witnesses in lexical ambiguity data, it is still not clear that quantum systems are *necessary* or would even provide

a computational advantage, in simulating the disambiguation process – further investigation will be needed in this regard.

Besides, the simulations conducted as part of the project were merely proof of principle, and the obtained results have yet to be compared with their classical analogues.

Investigating other parts-of-speech The next step should be to investigate the behaviour of other grammatical types (e.g. adjective, adverbs, ...) and more complex phrases and sentences, e.g. subject-verb-object sentences. However, the need for more psycholinguistics research may be a hindrance.

By doing so, we expect to provide a new compositional way of processing natural language data, which, although it comes from a different motivation, may be highly related to the approach of DisCoCat [42] or DisCoCirc [41] formalisms.

Improving the variational circuits In addition, even though we observe differences in data from words of different levels of ambiguity or different grammatical types, they are, in the empirical models, treated in the same way. This may particularly affect Chapter 4 simulations.

Examples of possible improvement may be allowing words to be represented as mixed states (i.e. probabilistic mixture of pure states) or pure states (i.e. superposition of states) – note that at the moment, nouns are only represented as pure states. Using the intuition of [148], we would expect homonymous nouns to be represented as mixed states and polysemous nouns as pure states.

Furthermore, the accuracy of the predictions and simulations may also increase by having an extra ancilla for verbs, thus allowing the verb to take in information from both the subject and object, even though one argument is not known. We could then represent underspecification by taking the partial trace over the system for which no information is provided. This representation for verbs would then be similar to the DisCoCat representation of a transitive verb, and by adopting this structure, we can train verb-states compatible with the DisCoCat formalism.

Including indefinite causal orders It is also quite clear that the process of disambiguating, even SV or VO phrases, is not entirely one-way (i.e., the probability distributions associated with the activation of subject and objects depend on the choice of verbs, as the verb provides context for the ambiguous nouns). Therefore,

to fully describe the disambiguation process, we must introduce the notion of *indefinite causal order*.

Indefinite causal orders have benefited from an increasing amount of research interest in the quantum foundations community, notably since a causal order can not only be probabilistic but also in *superposition* in (higher-order) processes such as the quantum switch.

The first question on the linguistic side is whether the disambiguation process is *causally separable* (i.e. correspond to the probabilistic mixture of causal orders) or *causally inseparable* (i.e. correspond to the superposition of causal orders). In the latter's case, this would provide an additional (and possibly more interpretable) advantage in using quantum resources. We could study this by calculating the so-called *causal separability fraction* introduced in [80].

Syntactic ambiguity

Further investigate the properties of our model Regarding our syntactic model, our line of research offers excellent promises relating to modelling cognitive processes using sheaves and presheaves. There are still many avenues to explore, e.g., the nature of the correlations between SF and difficulty or the model's applicability to a broader class of sentences. In addition, our model still underestimates the garden-path effects of both NP/S and NP/Z sentences. It is, therefore, imperative to identify the reason for this discrepancy, i.e. whether it be because of the choice of regression or due to a more fundamental factor such as backtracking. Finally, the uncertainty introduced from averaging data in the psycholinguistic datasets used widely hindered our preliminary results. Using different and more detailed datasets is a way to address this point.

Introducing an edit distance One possible criticism of our framework is that the different parses are treated as completely unrelated. In reality, this is not the case, as transformations between certain parses may be easy or hard. For instance, moving the head of a determiner by one place should be easier than changing the head of the whole sentence. These transformations between parses can also occur at different levels, for instance, within the same sentence fragment or across different fragments. Defining a measure of discrepancy between probability distributions over parses that considers this "transformation difficulty" is left as future work.

Studying reanalysis In addition, the main hypothesis behind garden-path effect underestimation, in surprisal theory at least, corresponds to backtracking or other related non-incremental processes. Contrary to surprisal, however, it would be fairly easy to alter our current parsing model to study non-incremental processes. In particular, this could be done by extending the morphisms from prefix order to standard inclusions and changing the choice of cover. This would amount to changing the causal order of interest. By comparing the causal fractions associated with different causal orders, we expect that the one(s) with the highest causal fraction would show up in eye-tracking data as the trajectories adopted by the different readers. In addition, it is not clear that only one causal order would be more advantageous as compared to others; in fact, we would expect multiple causal orders to have comparably high causal fractions. Hence, we would expect that the different causal fractions might predict which *reanalysis patterns* could be employed by readers and at which frequency each of the possible patterns is adopted.

BIBLIOGRAPHY

- [1] *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain, July 2004. Association for Computational Linguistics.
- [2] Samson Abramsky and Rui Soares Barbosa. The logic of contextuality. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021.
- [3] Samson Abramsky, Rui Soares Barbosa, Martti Karvonen, and Shane Mansfield. A comonadic view of simulation and quantum resources. volume 2019-June, 2019.
- [4] Samson Abramsky, Rui Soares Barbosa, and Shane Mansfield. Contextual fraction as a measure of contextuality. *Physical Review Letters*, 119, 2017.
- [5] Samson Abramsky, Rui Soares Barbosa, and Amy Searle. Combining contextuality and causality: a game semantics approach, 2023.
- [6] Samson Abramsky and Adam Brandenburger. The sheaf-theoretic structure of non-locality and contextuality. *New J. Phys.*, 13:113036, 2011.
- [7] Samson Abramsky and Bob Coecke. A categorical semantics of quantum protocols. volume 19, 2004.
- [8] Samson Abramsky and Mehrnoosh Sadrzadeh. Semantic unification: A sheaf theoretic approach to natural language. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8222, 2014.

- [9] Eneko Agirre, Lluís Màrquez, and Richard Wicentowski. Computational semantic analysis of language: Semeval-2007 and beyond. *Language Resources and Evaluation*, 43(2):97–104, Jun 2009.
- [10] Johan Ahrens, Elias Amselem, Adan Cabello, and Mohamed Bourennane. Two fundamental experimental tests of nonclassicality with qutrits, 2013.
- [11] Kazimierz Ajdukiewicz. Die syntaktische konnexitat. *Studia philosophica*, pages 1–27, 1935.
- [12] Barbara Amaral. Resource theory of contextuality. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 377(2157):20190010, 2019.
- [13] Mateus Araujo, Marco Túlio Quintino, Costantino Budroni, Marcelo Terra Cunha, and Adán Cabello. All noncontextuality inequalities for the n -cycle scenario. *Physical Review A*, 88(2), August 2013.
- [14] Suhas Arehalli, Brian Dillon, and Tal Linzen. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.
- [15] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C. Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando G.S.L. Brandao, David A. Buell, Brian Burkett, Yu Chen, Zijun Chen, Ben Chiaro, Roberto Collins, William Courtney, Andrew Dunsworth, Edward Farhi, Brooks Foxen, Austin Fowler, Craig Gidney, Marissa Giustina, Rob Graff, Keith Guerin, Steve Habegger, Matthew P. Harrigan, Michael J. Hartmann, Alan Ho, Markus Hoffmann, Trent Huang, Travis S. Humble, Sergei V. Isakov, Evan Jeffrey, Zhang Jiang, Dvir Kafri, Kostyantyn Kechedzhi, Julian Kelly, Paul V. Klimov, Sergey Knysh, Alexander Korotkov, Fedor Kostritsa, David Landhuis, Mike Lindmark, Erik Lucero, Dmitry Lyakh, Salvatore Mandrà, Jarrod R. McClean, Matthew McEwen, Anthony Megrant, Xiao Mi, Kristel Michielsen, Masoud Mohseni, Josh Mutus, Ofer Naaman, Matthew Neeley, Charles Neill, Murphy Yuezhen Niu, Eric Ostby, Andre Petukhov, John C. Platt, Chris Quintana,

- Eleanor G. Rieffel, Pedram Roushan, Nicholas C. Rubin, Daniel Sank, Kevin J. Satzinger, Vadim Smelyanskiy, Kevin J. Sung, Matthew D. Trevithick, Amit Vainsencher, Benjamin Villalonga, Theodore White, Z. Jamie Yao, Ping Yeh, Adam Zalcman, Hartmut Neven, and John M. Martinis. Quantum supremacy using a programmable superconducting processor. *Nature*, 574, 2019.
- [16] Ryan Babbush, Dominic W Berry, Robin Kothari, Rolando D Somma, and Nathan Wiebe. Exponential quantum speedup in simulating coupled classical oscillators. *arXiv preprint arXiv:2303.13012*, 2023.
- [17] John C. Baez and Aaron D. Lauda. *A Prehistory of n-Categorical Physics*, page 13–128. Cambridge University Press, 2011.
- [18] Satanjeev Banerjee and Ted Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. volume 2276, 2002.
- [19] Yehoshua Bar-Hillel and Rudolf Carnap. Semantic information. *British Journal for the Philosophy of Science*, 4:147–157, 1953.
- [20] Rui Soares Barbosa. On monogamy of non-locality and macroscopic averages: examples and preliminary results. *Electronic Proceedings in Theoretical Computer Science*, 172:36–55, dec 2014.
- [21] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. 2014.
- [22] John S. Bell. On the problem of hidden variables in quantum mechanics. *Reviews of Modern Physics*, 38, 1966.
- [23] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. volume 3, 2003.
- [24] Dominic W Berry, Graeme Ahokas, Richard Cleve, and Barry C Sanders. Efficient quantum algorithms for simulating sparse hamiltonians. *Communications in Mathematical Physics*, 270:359–371, 2007.
- [25] Thomas G Bever. The cognitive basis for linguistic structures. *Cognition and the development of language*, 1970.

- [26] John Neil Bohannon, Barbara Landau, and Lila Gleitman. Language and experience: Evidence from the blind child. *Language*, 62, 1986.
- [27] David Bohm. A suggested interpretation of the quantum theory in terms of "hidden" variables. i. *Phys. Rev.*, 85:166–179, Jan 1952.
- [28] David Bohm. A suggested interpretation of the quantum theory in terms of "hidden" variables. ii. *Phys. Rev.*, 85:180–193, Jan 1952.
- [29] Niels Bohr. On The Notions of Causality and Complementarity. In Jørgen Kalckar, editor, *Foundations of Quantum Physics II (1933–1958)*, volume 7 of *Niels Bohr Collected Works*, pages 325–338. Elsevier, 1996.
- [30] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. Training algorithm for optimal margin classifiers. 1992.
- [31] Claudio Delli Bovi, Luca Telesca, and Roberto Navigli. Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*, 3, 2015.
- [32] Peter D. Bruza, Kirsty Kitto, Brentyn J. Ramm, and Laurianne Sitbon. A probabilistic framework for analysing the compositionality of conceptual combinations. *Journal of Mathematical Psychology*, 67, 2015.
- [33] Adan Cabello, Simone Severini, and Andreas Winter. (non-)contextuality of physical theories as an axiom, 2010.
- [34] Adán Cabello, José M. Estebarez, and Guillermo García-Alcaine. Bell-kochen-specker theorem: A proof with 18 vectors. *Physics Letters, Section A: General, Atomic and Solid State Physics*, 212, 1996.
- [35] Henri Cartan. Idéaux et modules de fonctions analytiques de variables complexes. *Bulletin de la Société mathématique de France*, 78:29–64, 1950.
- [36] Eric G. Cavalcanti. Classical causal models for bell and kochen-specker inequality violations require fine-tuning. *Physical Review X*, 8, 2018.
- [37] Carlton M. Caves, Christopher A. Fuchs, and Rüdiger Schack. Quantum probabilities as bayesian probabilities. *Physical Review A*, 65(2), January 2002.

- [38] Jinho D Choi, Joel Tetreault, and Amanda Stent. It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 387–396, 2015.
- [39] N. Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124, Sep. 1956.
- [40] John F. Clauser, Michael A. Horne, Abner Shimony, and Richard A. Holt. Proposed experiment to test local hidden-variable theories. *Physical Review Letters*, 23, 1969.
- [41] Bob Coecke. *The Mathematics of Text Structure*, volume 20. 2021.
- [42] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning, 2010.
- [43] BNC Consortium. The british national corpus. Oxford Text Archive, 2007. XML Edition.
- [44] Justin Curry. Topological data analysis and cosheaves, 2015.
- [45] Pierre Deligne. La conjecture de Weil : I. *Publications Mathématiques de l’IHÉS*, 43:273–307, 1974.
- [46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [47] Stephen Dopkins, Robin K. Morris, and Keith Rayner. Lexical ambiguity and eye fixations in reading: A test of competing models of lexical ambiguity resolution. *Journal of Memory and Language*, 31, 1992.
- [48] Cristhiano Duarte and Barbara Amaral. Resource theory of contextuality for arbitrary prepare-and-measure experiments. *Journal of Mathematical Physics*, 59, 2018.

- [49] E. N. Dzhafarov, Ru Zhang, and Janne Kujala. Is there contextuality in behavioural and social systems? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374, 2016.
- [50] Ehtibar N. Dzhafarov. The contextuality-by-default view of the sheaf-theoretic approach to contextuality, 2023.
- [51] Ehtibar N. Dzhafarov, Víctor H. Cervantes, and Janne V. Kujala. Contextuality in canonical systems of random variables. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 375, 2017.
- [52] Ehtibar N. Dzhafarov and Janne V. Kujala. Context–content systems of random variables: The contextuality-by-default theory. *Journal of Mathematical Psychology*, 74, 2016.
- [53] Ehtibar N. Dzhafarov, Janne V. Kujala, and Jan Åke Larsson. Contextuality in three types of quantum-mechanical systems. *Foundations of Physics*, 45, 2015.
- [54] Philip Edmonds. Senseval: The evaluation of word sense disambiguation systems. 2002.
- [55] Philip Edmonds and Scott Cotton. Senseval-2: Overview. 2001.
- [56] Susan F. Ehrlich and Keith Rayner. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655, 1981.
- [57] Samuel Eilenberg and Saunders MacLane. General theory of natural equivalences. *Transactions of the American Mathematical Society*, 58, 1945.
- [58] A. Einstein, B. Podolsky, and N. Rosen. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.*, 47:777–780, May 1935.
- [59] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm, 2014.
- [60] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm applied to a bounded occurrence constraint problem, 2015.

- [61] Edward Farhi and Hartmut Neven. Classification with Quantum Neural Networks on Near Term Processors, 2018.
- [62] Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. Introducing and evaluating ukwac , a very large web-derived corpus of english. 2008.
- [63] Arthur Fine. Hidden variables, joint probability, and the bell inequalities. *Phys. Rev. Lett.*, 48:291–295, Feb 1982.
- [64] J.R. Firth. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis: Special Volume of the Philological Society*, 1957.
- [65] Fulvio Flamini, Arne Hamann, Sofiène Jerbi, Lea M. Trenkwalder, Hendrik Poulsen Nautrup, and Hans J. Briegel. Photonic architecture for reinforcement learning. *New Journal of Physics*, 22, 2020.
- [66] Brendan Fong and David I Spivak. Seven sketches in compositionality: An invitation to applied category theory, 2018.
- [67] Stefan L. Frank. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, 5(3):475–494, 2013.
- [68] Lyn Frazier. *Sentence processing: A tutorial review.*, pages 559–586. Attention and performance 12: The psychology of reading. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1987.
- [69] Lyn Frazier and Keith Rayner. Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, 29, 1990.
- [70] Peter Freyd. The axiom of choice. *Journal of Pure and Applied Algebra*, 19:103–125, 1980.
- [71] Steven Frisson. Semantic underspecification in language processing. *Linguistics and Language Compass*, 3, 2009.
- [72] Steven Frisson and Martin J. Pickering. The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning Memory and Cognition*, 25, 1999.

- [73] Susan M. Garnsey, Neal J. Pearlmutter, Elizabeth Myers, and Melanie A. Lockett. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37(1):58–93, 1997.
- [74] D Gentner. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Language development: Vol. 2. Language, thought, and culture*, 2, 1982.
- [75] Dedre Gentner. Some interesting differences. *Cognition and brain theory*, 4, 1981.
- [76] Dedre Gentner and Ilene M. France. *The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs*. 2013.
- [77] Edward Gibson and Neal J Pearlmutter. Distinguishing serial and parallel parsing. *Journal of Psycholinguistic Research*, 29:231–240, 2000.
- [78] Marissa Giustina, Marijn AM Versteegh, Sören Wengerowsky, Johannes Handsteiner, Armin Hochrainer, Kevin Phelan, Fabian Steinlechner, Johannes Kofler, Jan-Åke Larsson, Carlos Abellán, et al. Significant-loophole-free test of bell’s theorem with entangled photons. *Physical review letters*, 115(25):250401, 2015.
- [79] Stefano Gogioso and Nicola Pinzani. The Sheaf-Theoretic Structure of Definite Causality. *Electronic Proceedings in Theoretical Computer Science*, 343:301–324, Sep 2021.
- [80] Stefano Gogioso and Nicola Pinzani. *The geometry of causality*, 2023.
- [81] Robert Goldblatt. *Topoi: The Categorical Analysis of Logic*. Dover Publications, 1983.
- [82] Daniel Gottesman. *The heisenberg representation of quantum computers*, 1998.
- [83] Daniel Grodner, Edward Gibson, Vered Argaman, and Maria Babyonyshev. Against repair-based reanalysis in sentence comprehension. *Journal of Psycholinguistic Research*, 32:141–166, 2003.

- [84] Alexander Grothendieck. Formule de lefschetz et rationalité des fonctions ζ . 1966.
- [85] Alexandre Grothendieck. Sur quelques points d'algèbre homologique. *Tohoku Mathematical Journal, Second Series*, 9(2):119–183, 1957.
- [86] Lov K. Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '96, page 212–219, New York, NY, USA, 1996. Association for Computing Machinery.
- [87] John Hale. A probabilistic Earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*, 2001.
- [88] John Hale. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123, Mar 2003.
- [89] John Hale. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672, 2006.
- [90] Zellig S. Harris. Distributional structure. *WORD*, 10, 1954.
- [91] Bas Hensen, Hannes Bernien, Anaïs E Dréau, Andreas Reiserer, Norbert Kalb, Machiel S Blok, Just Ruitenbergh, Raymond FL Vermeulen, Raymond N Schouten, Carlos Abellán, et al. Loophole-free bell inequality violation using electron spins separated by 1.3 kilometres. *Nature*, 526(7575):682–686, 2015.
- [92] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9, 1997.
- [93] V. M. Holmes. *Syntactic parsing: In search of the garden path.*, pages 587–599. Attention and performance 12: The psychology of reading. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1987.
- [94] Mark Howard, Joel Wallman, Victor Veitch, and Joseph Emerson. Contextuality supplies the 'magic' for quantum computation. *Nature*, 510, 2014.

- [95] Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. Surprisal does not explain syntactic disambiguation difficulty: evidence from a large-scale benchmark. 2023.
- [96] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Embeddings for word sense disambiguation: An evaluation study. volume 2, 2016.
- [97] Bart Jacobs, Aleks Kissinger, and Fabio Zanasi. Causal inference by string diagram surgery, 2019.
- [98] Peter T Johnstone. *Sketches of an Elephant: A Topos Theory Compendium*, volume 2. Oxford University Press, 2002.
- [99] Matt Jones. Relating causal and probabilistic approaches to contextuality. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 377, 2019.
- [100] Martin Joos. Description of language design. *Journal of the Acoustical Society of America*, 22, 1950.
- [101] Daniel Jurafsky. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20(2):137–194, 1996.
- [102] Daniel Jurafsky and James H Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*.
- [103] A. Kilgarriff and J. Rosenzweig. Framework and results for english senseval. *Language Resources and Evaluation*, 34, 2000.
- [104] Aleks Kissinger, Matty Hoban, and Bob Coecke. Equivalence of relativistic causal structure and process terminality, 2017.
- [105] Aleks Kissinger and Sander Uijlen. A categorical semantics for causal structure. *Logical Methods in Computer Science*, 15, 2019.
- [106] Alexander A. Klyachko, M. Ali Can, Sinem Binicioğlu, and Alexander S. Shumovsky. Simple test for hidden variables in spin-1 systems. *Physical Review Letters*, 101(2), July 2008.

- [107] Simon Kochen and E. Specker. The problem of hidden variables in quantum mechanics. *Indiana University Mathematics Journal*, 17, 1967.
- [108] Saul A. Kripke. Semantical analysis of intuitionistic logic i. In J.N. Crossley and M.A.E. Dummett, editors, *Formal Systems and Recursive Functions*, volume 40 of *Studies in Logic and the Foundations of Mathematics*, pages 92–130. Elsevier, 1965.
- [109] Janne V. Kujala and Ehtibar N. Dzhafarov. Proof of a conjecture on contextuality in cyclic systems with binary variables. *Foundations of Physics*, 46, 2016.
- [110] Janne V. Kujala and Ehtibar N. Dzhafarov. Measures of contextuality and non-contextuality. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 377, 2019.
- [111] Janne V. Kujala, Ehtibar N. Dzhafarov, and Jan Åke Larsson. Necessary and sufficient conditions for an extended noncontextuality in a broad class of quantum mechanical systems. *Physical Review Letters*, 115, 2015.
- [112] Joachim Lambek. The mathematics of sentence structure. *The American Mathematical Monthly*, 65(3):154–170, 1958.
- [113] Joachim Lambek. From lambda-calculus to cartesian closed categories. *To HB Curry: essays on combinatory logic, lambda calculus and formalism*, pages 375–402, 1980.
- [114] Joachim Lambek and Philip J Scott. *Introduction to higher-order categorical logic*, volume 7. Cambridge University Press, 1988.
- [115] F William Lawvere. An elementary theory of the category of sets. *Proceedings of the national academy of sciences*, 52(6):1506–1511, 1964.
- [116] F William Lawvere. Quantifiers and sheaves. In *Actes du congrès international des mathématiciens, Nice*, volume 1, pages 329–334, 1970.
- [117] Yoong Keok Lee and Hwee Tou Ng. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. 2002.
- [118] Jean Leray. Sur la forme des espaces topologiques et sur les points fixes des représentations. *Journal de Mathématiques Pures et Appliquées*, 24:95–167, 1945.

- [119] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries. 1986.
- [120] Kin Ian Lo, Mehrnoosh Sadrzadeh, and Shane Mansfield. A model of anaphoric ambiguities using sheaf theoretic quantum-like contextuality and bert. volume 366, 2022.
- [121] Kin Ian Lo, Mehrnoosh Sadrzadeh, and Shane Mansfield. Generalised winograd schema and its contextuality. *Electronic Proceedings in Theoretical Computer Science*, 384, 2023.
- [122] Robin Lorenz, Anna Pearson, Konstantinos Meichanetzidis, Dimitri Kartsaklis, and Bob Coecke. QNLP in Practice: Running Compositional Models of Meaning on a Quantum Computer, 2021.
- [123] Robin Lorenz and Sean Tull. Causal models in string diagrams, 2023.
- [124] Daniel Loureiro and Alípio Mário Jorge. Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. 2020.
- [125] Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. 2018.
- [126] Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. Incorporating glosses into neural word sense disambiguation. volume 1, 2018.
- [127] Saunders Mac Lane. *Categories for the working mathematician*, volume 5. Springer Science & Business Media, 2013.
- [128] Saunders MacLane and Ieke Moerdijk. *Sheaves in geometry and logic: A first introduction to topos theory*. Springer Science & Business Media, 2012.
- [129] Lars S. Madsen, Fabian Laudenbach, Mohsen Falamarzi Askarani, Fabien Rortais, Trevor Vincent, Jacob F.F. Bulmer, Filippo M. Miatto, Leonhard Neuhaus, Lukas G. Helt, Matthew J. Collins, Adriana E. Lita, Thomas Gerrits, Sae Woo Nam, Varun D. Vaidya, Matteo Menotti, Ish Dhand, Zachary Vernon, Nicolás Quesada, and Jonathan Lavoie. Quantum computational advantage with a programmable photonic processor. *Nature*, 606, 2022.

- [130] Shane Mansfield and Elham Kashefi. Quantum advantage from sequential-transformation contextuality. *Physical Review Letters*, 121(23), dec 2018.
- [131] Oren Melamud, Jacob Goldberger, and Ido Dagan. context2vec: Learning generic context embedding with bidirectional lstm. 2016.
- [132] N. David Mermin. Simple unified form for the major no-hidden-variables theorems. *Physical Review Letters*, 65, 1990.
- [133] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. 2013.
- [134] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov 1995.
- [135] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.
- [136] George A. Miller, Claudia Leacock, Randee Teng, and Ross T. Bunker. A semantic concordance. 1993.
- [137] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii. Quantum circuit learning. *Phys. Rev. A*, 98:032309, Sep 2018.
- [138] D. C. Mitchell. *Lexical guidance in human parsing: Locus and processing characteristics.*, pages 601–618. Attention and performance 12: The psychology of reading. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1987.
- [139] Allison Mullaly, Christina Gagné, Thomas Spalding, and Kristan Marchak. Examining ambiguous adjectives in adjective-noun phrases: Evidence for representation as a shared core-meaning with sense specialization. *The Mental Lexicon*, 5:87–114, 06 2010.
- [140] Maarten Van Den Nest. Universal quantum computation with little entanglement. *Physical Review Letters*, 110, 2013.
- [141] Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010.

- [142] Judea Pearl. *Causality: Models, reasoning, and inference, second edition*. 2011.
- [143] A. Peres. Two simple proofs of the kochen-specker theorem. *Journal of Physics A: General Physics*, 24, 1991.
- [144] Asher Peres. Separability criterion for density matrices. *Phys. Rev. Lett.*, 77:1413–1415, Aug 1996.
- [145] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. volume 1, 2018.
- [146] Martin J. Pickering and Steven Frisson. Processing ambiguous verbs: Evidence from eye movements. *Journal of Experimental Psychology: Learning Memory and Cognition*, 27, 2001.
- [147] Martin J Pickering and Matthew J Traxler. Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(4):940, 1998.
- [148] Robin Piedeleu, Dimitri Kartsaklis, Bob Coecke, and Mehrnoosh Sadrzadeh. Open system categorical quantum semantics in natural language processing. volume 35, 2015.
- [149] Robin Piedeleu and Fabio Zanasi. An introduction to string diagrams for computer scientists, 2023.
- [150] Martin B. Plenio and Vlatko Vedral. Teleportation, entanglement and thermodynamics in the quantum world. *Contemporary Physics*, 39(6):431–446, nov 1998.
- [151] Grusha Prasad and Tal Linzen. Rapid syntactic adaptation in self-paced reading: Detectable, but only with many participants. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(7):1156, 2021.
- [152] Bradley L Pritchett. *Grammatical competence and parsing performance*. University of Chicago Press, 1992.
- [153] Qiskit contributors. Qiskit: An open-source framework for quantum computing, 2023.

- [154] Ganesh Ramakrishnan, Apurva Jadhav, Ashutosh Joshi, Soumen Chakrabarti, and Pushpak Bhattacharyya. Question answering via bayesian inference on lexical relations. 2003.
- [155] Keith Rayner. Visual attention in reading: Eye movements reflect cognitive processes. *Memory & Cognition*, 5, 1977.
- [156] Keith Rayner and Susan A. Duffy. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14, 1986.
- [157] Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. Quantum support vector machine for big data classification. *Physical Review Letters*, 113, 2014.
- [158] Jane J. Robinson. Dependency structures and transformational rules. *Language*, 46(2):259–285, 1970.
- [159] Wenjamin Rosenfeld, Daniel Burchardt, Robert Garthoff, Kai Redeker, Norbert Ortengel, Markus Rau, and Harald Weinfurter. Event-ready bell test using entangled atoms simultaneously closing detection and locality loopholes. *Physical review letters*, 119(1):010402, 2017.
- [160] Maria Schuld, Alex Bocharov, Krysta M. Svore, and Nathan Wiebe. Circuit-centric quantum classifiers. *Physical Review A*, 101(3), mar 2020.
- [161] Hinrich Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24, 1998.
- [162] Peter Selinger. Dagger compact closed categories and completely positive maps. *Electronic Notes in Theoretical Computer Science*, 170:139–163, 03 2007.
- [163] Jean-Pierre Serre. Faisceaux algébriques cohérents. *Annals of Mathematics*, pages 197–278, 1955.
- [164] Lynden K Shalm, Evan Meyer-Scott, Bradley G Christensen, Peter Bierhorst, Michael A Wayne, Martin J Stevens, Thomas Gerrits, Scott Glancy, Deny R Hamel, Michael S Allman, et al. Strong loophole-free test of local realism. *Physical review letters*, 115(25):250402, 2015.

- [165] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.
- [166] Peter W Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM review*, 41(2):303–332, 1999.
- [167] Ekaterina Shutova. Automatic metaphor interpretation as a paraphrasing task. 2010.
- [168] Nathaniel J Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319, 2013.
- [169] A. M. Steane. A quantum computer only needs one universe. *Studies in History and Philosophy of Science Part B - Studies in History and Philosophy of Modern Physics*, 34, 2003.
- [170] Mark Steedman. *Combinators and Grammars*, pages 417–442. Springer Netherlands, Dordrecht, 1988.
- [171] W. Forrest Stinespring. Positive functions on c^* -algebras. *Proceedings of the American Mathematical Society*, 6, 1955.
- [172] Simon Storz, Josua Schär, Anatoly Kulikov, Paul Magnard, Philipp Kurpiers, Janis Lütolf, Theo Walter, Adrian Copetudo, Kevin Reuer, Abdulkadir Akin, Jean-Claude Besse, Mihai Gabureac, Graham J. Norris, Andrés Rosario, Ferran Martin, José Martinez, Waldimar Amaya, Morgan W. Mitchell, Carlos Abellan, Jean-Daniel Bancal, Nicolas Sangouard, Baptiste Royer, Alexandre Blais, and Andreas Wallraff. Loophole-free bell inequality violation with superconducting circuits. *Nature*, 617(7960):265–270, May 2023.
- [173] Patrick Sturt, Martin J Pickering, and Matthew W Crocker. Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, 40(1):136–150, 1999.
- [174] Michael K. Tanenhaus, James M. Leiman, and Mark S. Seidenberg. Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior*, 18, 1979.
- [175] Wilson L Taylor. “cloze procedure”: A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

- [176] M. Tierney. *Axiomatic Sheaf Theory : Some Constructions and Applications*, pages 249–326. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.
- [177] Myles Tierney. Sheaf theory and the continuum hypothesis. In F. W. Lawvere, editor, *Toposes, Algebraic Geometry and Logic*, pages 13–42, Berlin, Heidelberg, 1972. Springer Berlin Heidelberg.
- [178] Michael Tomasello, Nameera Akhtar, , Kelly Dodson, and Laura Rekau. Differential productivity in young children’s use of nouns and verbs. *Journal of Child Language*, 24, 1997.
- [179] Matthew J. Traxler, Martin J. Pickering, and Charles Clifton. Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, 39(4):558–592, 1998.
- [180] John C. Trueswell and Michael K. Tanenhaus. Tense, temporal context and syntactic ambiguity resolution. *Language and Cognitive Processes*, 6(4):303–338, 1991.
- [181] John C Trueswell, Michael K Tanenhaus, and Christopher Kello. Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental psychology: Learning, memory, and Cognition*, 19(3):528, 1993.
- [182] Carlo A. Trugenberger. Quantum pattern recognition. *Quantum Information Processing*, 1(6):471–493, Dec 2002.
- [183] Kim Vallée, Pierre-Emmanuel Emeriau, Boris Bourdoncle, Adel Sohbi, Shane Mansfield, and Damian Markham. Corrected bell and non-contextuality inequalities for realistic experiments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382(2268):20230011, 2024.
- [184] Roger P.G. van Gompel, Martin J. Pickering, Jamie Pearson, and Simon P. Liv-ersedge. Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, 52(2):284–307, 2005.
- [185] Marten Van Schijndel and Tal Linzen. Modeling garden path effects without explicit hierarchical syntax. In *CogSci*, 2018.

- [186] Marten van Schijndel and Tal Linzen. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988, 2021.
- [187] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [188] V. Vedral, M. B. Plenio, K. Jacobs, and P. L. Knight. Statistical inference, distinguishability of quantum states, and quantum entanglement. *Phys. Rev. A*, 56:4452–4455, Dec 1997.
- [189] N. N. Vorob'ev. Consistent families of measures and their extensions. *Theory of Probability & Its Applications*, 7(2):147–163, 1962.
- [190] Rafael Wagner, Roberto D. Baldijão, Alisson Tezzin, and Bárbara Amaral. Using a resource theoretic perspective to witness and engineer quantum generalized contextuality for prepare-and-measure scenarios, 2023.
- [191] Daphne Wang. The corpus dataset. https://github.com/wangdaphne/Cyclic_models, 2022.
- [192] Daphne Wang. Empirical models and signalling fractions of garden-path sentences. <https://github.com/wangdaphne/garden-path-SF-dataset>, 2023.
- [193] Daphne Wang and Mehrnoosh Sadrzadeh. Causality and signalling of garden-path sentences. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382(2268):20230013, 2024.
- [194] Daphne Wang, Mehrnoosh Sadrzadeh, Samson Abramsky, and Víctor H. Cervantes. On the quantum-like contextuality of ambiguous phrases. 2021.
- [195] Daphne Wang, Mehrnoosh Sadrzadeh, Samson Abramsky, H. Víctor, and Cervantes. Analysing ambiguous nouns and verbs with quantum contextuality tools. *Journal of Cognitive Science*, 22, 2021.

- [196] Zheng Wang and Jerome R. Busemeyer. A quantum question order model supported by empirical tests of an a priori and precise prediction. *Topics in Cognitive Science*, 5, 2013.
- [197] Warren Weaver. Translation. In *Proceedings of the Conference on Mechanical Translation*, 1952.
- [198] Paul Wilson, Dan Ghica, and Fabio Zanasi. String diagrams for non-strict monoidal categories, 2022.
- [199] William K. Wootters. Entanglement of formation and concurrence. *Quantum Info. Comput.*, 1(1):27–44, jan 2001.
- [200] Pierre yves Vandenbussche, Tony Scerri, and Ron Daniel Jr. Word sense disambiguation with transformer models. *Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6)*, 2021.
- [201] William Zeng and Bob Coecke. Quantum algorithms for compositional natural language processing. volume 221, 2016.
- [202] Han-Sen Zhong, Hui Wang, Yu-Hao Deng, Ming-Cheng Chen, Li-Chao Peng, Yi-Han Luo, Jian Qin, Dian Wu, Xing Ding, Yi Hu, et al. Quantum computational advantage using photons. *Science*, 370(6523):1460–1463, 2020.
- [203] Zhi Zhong and Hwee Tou Ng. It makes sense: A wide-coverage word sense disambiguation system for free text. 2010.
- [204] D. Zhu, N. M. Linke, M. Benedetti, K. A. Landsman, N. H. Nguyen, C. H. Alderete, A. Perdomo-Ortiz, N. Korda, A. Garfoot, C. Brecque, L. Egan, O. Perdomo, and C. Monroe. Training of quantum circuits on a hybrid quantum computer. *Science Advances*, 5(10):eaaw9918, 2019.

Part IV

APPENDIX

Chapter A

SECTIONS OF A PRESHEAF AND THE SHEAFICATION OF A PRESHEAF

Bundles A *bundle* is an alternative presentation of families of sets $\{X_i\}_{i \in I}$ as a map $p : E \rightarrow I$, where $E = \bigsqcup_{i \in I} X_i = \{(i, x) | x \in X_i, i \in I\}$, and each of the X_i is a set. The map p in this case is simply defined as:

$$p :: (i, x) \mapsto i \tag{A.1}$$

The space E is referred to as the *espace étalé* or the *total space*, while I is referred to as the *base space* of the bundle. The equivalence of between families of sets and the bundle can be seen as each of the X_i can be retrieved from the map p as:

$$X_i = p^{-1}(i) \tag{A.2}$$

It is usually said that the set X_i “sits on top” of the point i in the base space. The sets X_i are called the *stalks* or *fibres* of p at $i \in I$ and each of the elements of a given X_i are called the *germs* for the stalk X_i . This terminology comes from a vegetal analogy where, stalks of a plant, e.g. say wheat, grows up from the soil (here the base space), each each of the stalk consists germs (see Fig. A.1). We will also define a *section* of

the bundle p as a map $s : I \rightarrow E$ s.t. $p \circ s = id_I$; the idea is that sections will select a single germ in each of the fibres of p .

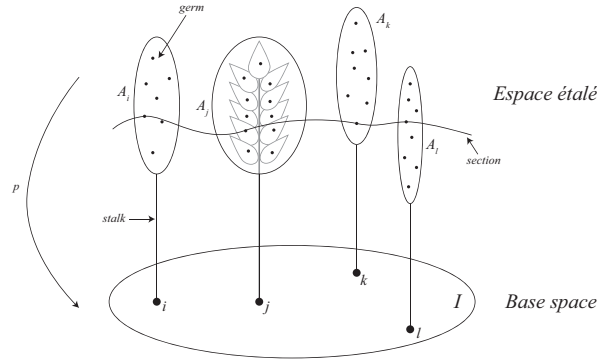


Figure A.1: Illustration of a bundle.

Furthermore, for every function $p : E \rightarrow I$ (for any sets E and I) defines a family of sets $\{X_i\}_{i \in I}$ where the X_i 's are obtained exactly as $X_i = p^{-1}(i)$.

So far, no structure is assumed on the base and étalé space, and notably all of the points of I are considered as unrelated. Let us now consider the case when E and I are topological spaces. Then, if $p : E \rightarrow I$ is a continuous map, then the associated bundle will have some nice properties as well. In particular, the continuity condition (with respect to the topology of E) implies that stalks are “glued together” in the sense that any two open neighbourhoods of a point $x \in E$ will be mapped to (open) sets in I which contains the point $p(x)$. The notion of section is then extend as follows. Given a continuous map $p : E \rightarrow I$, we define a section of the open set $U \subseteq I$ as a continuous map $s : U \rightarrow E$ s.t. the following pullback square commutes in **Top**.

$$\begin{array}{ccc}
 p^{-1}U & \hookrightarrow & E \\
 \downarrow pU & \lrcorner & \downarrow p \\
 U & \hookrightarrow & I
 \end{array}
 \quad (A.3)$$

In turn, these sections give rise to the *sheaf of sections* associated each any continuous bundle $p : E \rightarrow I$, as the functor:

$$\begin{array}{ccc}
 \Gamma_p : \mathcal{T}(I)^{op} & \rightarrow & \mathbf{Sets} \\
 U & \mapsto & \left\{ s : U \rightarrow E \mid U \xrightarrow{s} E \xrightarrow{p} I = U \hookrightarrow X \right\}
 \end{array}
 \quad (A.4)$$

And the morphisms are defined as restriction morphisms as described in Section 1.1.2. By continuity of p , this will indeed satisfy the sheaf condition. In addition, given a continuous bundle p , the sections over U are indeed the elements of Γ_p .

Sheafification We have seen that the sections of a bundle indeed correspond to the elements of the images of a certain (pre)sheaf associated with the bundle. We now try to go in the reverse direction, namely, the elements of PU for an arbitrary $P : \mathcal{T}(X)^{op} \rightarrow \mathbf{Sets}$ will correspond to the sections of a bundle $p : E \rightarrow X$. As a bonus, the construction of p also gives us a construction of a sheaf from an arbitrary presheaf, which satisfy a universal property; this construction is therefore known as *sheafification*.

We start by defining the *germs* at a point $x \in U$ of an element $s \in PU$ as the following set:

$$\begin{aligned} germ_x s = \{ & t \in PV \mid V \text{ open neighbourhood of } x \\ & \wedge \exists W \text{ open neighbourhood of } x. s|_W = t|_W \} \end{aligned} \quad (\text{A.5})$$

We can then define the *stalks* of a presheaf P at $x \in X$ as:

$$P_x = \{ germ_x s \mid \exists U \text{ open neighbourhood of } x. s \in PU \} \quad (\text{A.6})$$

Then, by defining the set:

$$\Lambda_P = \bigsqcup_{x \in X} P_x \quad (\text{A.7})$$

we can define the following bundle:

$$\begin{aligned} p : \quad \Lambda_P & \rightarrow X \\ (x, germ_x s) & \mapsto x \end{aligned} \quad (\text{A.8})$$

Furthermore, for any $s \in PU$, we can define the following map:

$$\begin{aligned} \tilde{s} : U & \rightarrow \Lambda_P \\ x & \mapsto (x, germ_x s) \end{aligned} \quad (\text{A.9})$$

and it is not hard to verify that this is indeed a section of the bundle p . This bundle will, in turn, give rise to a sheaf of sections. This completes the sheafification process.

Chapter B

PROOF OF THE CHSH INEQUALITY

In order to prove (1.28), we start by obtaining a bound for the quantity $|\langle a'b \rangle - \langle a'b' \rangle|$. Since the hidden variable model should give back the observed probability distributions we have:

$$\langle a'b \rangle = \int_{\Lambda} d\lambda p(\lambda) A(a', \lambda) B(b, \lambda) \quad (\text{B.1})$$

where $A : \{a, a'\} \times \Lambda \rightarrow \{\pm 1\}$ and $B : \{b, b'\} \times \Lambda \rightarrow \{\pm 1\}$ are function associating a pair of input and hidden-variable with the deterministic outcome this environment gives out. Similarly, we have:

$$\langle a'b' \rangle = \int_{\Lambda} d\lambda p(\lambda) A(a', \lambda) B(b', \lambda) \quad (\text{B.2})$$

and:

$$|\langle a'b \rangle - \langle a'b' \rangle| = \left| \int_{\Lambda} d\lambda p(\lambda) (A(a', \lambda) B(b, \lambda) - A(a', \lambda) B(b', \lambda)) \right| \quad (\text{B.3})$$

Now, since the hidden variables determine the values of a, a', b, b' simultaneously, there is nothing stopping us from writing:

$$|\langle a'b \rangle - \langle a'b' \rangle| = \left| \int_{\Lambda} d\lambda p(\lambda) (A(a', \lambda)B(b, \lambda) - A(a', \lambda)B(b', \lambda)) \right. \\ \left. A(a, \lambda)B(b, \lambda)A(a', \lambda)B(b', \lambda) - A(a, \lambda)B(b, \lambda)A(a', \lambda)B(b', \lambda) \right| \quad (\text{B.4})$$

$$= \left| \int_{\Lambda} d\lambda p(\lambda) A(a', \lambda)B(b, \lambda) (1 + A(a, \lambda)B(b', \lambda)) \right. \\ \left. - \int_{\Lambda} d\lambda p(\lambda) A(a', \lambda)B(b', \lambda) (1 + A(a, \lambda)B(b, \lambda)) \right| \quad (\text{B.5})$$

$$= \left| \int_{\Lambda} d\lambda p(\lambda) A(a', \lambda)B(b, \lambda) (1 - A(a, \lambda)B(b', \lambda)) \right. \\ \left. - \int_{\Lambda} d\lambda p(\lambda) A(a', \lambda)B(b', \lambda) (1 - A(a, \lambda)B(b, \lambda)) \right| \quad (\text{B.6})$$

Focusing on (B.5), for now, we can apply the triangle inequality (twice) to obtain:

$$|\langle a'b \rangle - \langle a'b' \rangle| \leq \left| \int_{\Lambda} d\lambda p(\lambda) A(a', \lambda)B(b, \lambda) (1 + A(a, \lambda)B(b', \lambda)) \right| \\ + \left| \int_{\Lambda} d\lambda p(\lambda) A(a', \lambda)B(b', \lambda) (1 + A(a, \lambda)B(b, \lambda)) \right| \quad (\text{B.7})$$

$$\leq \int_{\Lambda} d\lambda \left| p(\lambda) A(a', \lambda)B(b, \lambda) (1 + A(a, \lambda)B(b', \lambda)) \right| \\ + \int_{\Lambda} d\lambda \left| p(\lambda) A(a', \lambda)B(b', \lambda) (1 + A(a, \lambda)B(b, \lambda)) \right| \quad (\text{B.8})$$

Now, since $A(a, \lambda), A(a', \lambda), B(b, \lambda), B(b', \lambda) \in \{\pm 1\}$ for all $\lambda \in \Lambda$, then:

$$|A(a', \lambda)B(b, \lambda)| = |A(a', \lambda)B(b', \lambda)| = 1 \quad (\text{B.9})$$

And:

$$p(\lambda) \geq 0 \quad (\text{B.10})$$

$$1 \pm A(a, \lambda)B(b, \lambda), 1 \pm A(a, \lambda)B(b', \lambda) \geq 0 \quad (\text{B.11})$$

So, (B.8) becomes:

$$|\langle a'b \rangle - \langle a'b' \rangle| \leq \int_{\Lambda} d\lambda p(\lambda) (1 + A(a, \lambda)B(b', \lambda)) + \int_{\Lambda} d\lambda p(\lambda) (1 + A(a, \lambda)B(b, \lambda)) \quad (\text{B.12})$$

Now, using:

$$\int_{\Lambda} d\lambda p(\lambda) = 1 \quad (\text{B.13})$$

we get:

$$|\langle a'b \rangle - \langle a'b' \rangle| \leq 2 + \int_{\Lambda} d\lambda p(\lambda) A(a, \lambda)B(b', \lambda) + \int_{\Lambda} d\lambda p(\lambda) A(a, \lambda)B(b, \lambda) = 2 + \langle ab' \rangle + \langle ab \rangle \quad (\text{B.14})$$

Similarly, starting from (B.6), we can adopt a similar reasoning to get:

$$|\langle a'b \rangle - \langle a'b' \rangle| \leq 2 - (\langle ab' \rangle + \langle ab \rangle) \quad (\text{B.15})$$

Now, using both (B.14) and (B.15), this gives:

$$|\langle a'b \rangle - \langle a'b' \rangle| \leq 2 - |\langle ab' \rangle + \langle ab \rangle| \quad (\text{B.16})$$

which by rearranging gives the CHSH equation:

$$|\langle ab' \rangle + \langle ab \rangle + \langle a'b \rangle - \langle a'b' \rangle| \leq 2 \quad (\text{B.17})$$

Chapter C

ORIGINAL PROOFS

C.1 Proof of proposition 1.44

In order to prove Proposition 1.44 we need some results from CbD and M-contextuality. In the CbD framework, given a cyclic system, or more generally a system for which every content is part of exactly 2 contexts, we want to minimise the probability $P[S_q^c = S_q^{c'}] = \sum_{o \in O} P[S_q^c = S_q^{c'} = o]$ (where O is the set of possible outcomes) for a globally imposed joint distribution S across all contexts (coupling), which agrees with the observed distributions.

Lemma C.1. *Given a content q and contexts c, c' containing q and outcome o , the maximum of $P[S_q^c = S_q^{c'} = o]$ for any coupling of the system is given by :*

$$\min \left(P[R_q^c = o], P[R_q^{c'} = o] \right) \tag{C.1}$$

Proof. We need a coupling to be compatible with the observed probability distributions, i.e. that the marginals of S coincide with the original distributions. This condition means that:

$$\sum_{o' \in O} P[S_q^c = o, S_q^{c'} = o'] = P[R_q^c = o] \tag{C.2}$$

for each context c, c' sharing the content q , and for every value $o \in O$. In particular, this implies both of the following inequalities:

$$P \left[S_q^c = o, S_q^{c'} = o \right] \leq P \left[R_q^c = o \right] \quad (\text{C.3})$$

$$P \left[S_q^c = o, S_q^{c'} = o \right] \leq P \left[R_q^{c'} = o \right] \quad (\text{C.4})$$

and so:

$$P \left[S_q^c = o, S_q^{c'} = o \right] \leq \min \left(P \left[R_q^c = o \right], P \left[R_q^{c'} = o \right] \right) \quad (\text{C.5})$$

In addition, given any system with content q , it is always possible to construct a coupling for which $P \left[S_q^c = S_q^{c'} \right]$ does attain its maximum (Theorem 3.3 of [52]). The above bound is therefore saturated. \square

One consequence of this is that:

$$\min P \left[S_q^c \neq S_q^{c'} \right] = 1 - \max P \left[S_q^c = S_q^{c'} \right] \quad (\text{C.6})$$

We now use one of the main results about the correspondence between CbD and M-contextuality.

Proposition C.2 (Proposition 8.4 of [99]). *Given a measurement system (i.e. context-content system with associated probability distributions), for each compatible canonical model \mathcal{M} , there exists a coupling S such that:*

$$\Delta_{c,c'}(F_q) = P \left[S_q^c \neq S_q^{c'} \right] \quad (\text{C.7})$$

for every content q . Conversely, for every coupling S , there exists a canonical model \mathcal{M} such that (C.7) is satisfied.

Corollary C.3. *The minimum of direct influence given a content q and pair of contexts c, c' , coincides with the minimum for $P \left[S_q^c \neq S_q^{c'} \right]$.*

We can now prove Proposition 1.

Proof of Proposition 1.44. By definition, we have:

$$\Delta = \sum_q \left| \langle R_q^{c_q} \rangle - \langle R_q^{c'_q} \rangle \right| \quad (\text{C.8})$$

Since only binary variables are considered for this definition to make sense, each individual term of the sum is given by:

$$\begin{aligned} \left| \langle R_q^{c_q} \rangle - \langle R_q^{c'_q} \rangle \right| &= \left| P [R_q^{c_q} = +1] - P [R_q^{c_q} = -1] - P [R_q^{c'_q} = +1] + P [R_q^{c'_q} = -1] \right| \\ &= 2 \left| P [R_q^{c_q} = +1] - P [R_q^{c'_q} = +1] \right| \end{aligned} \quad (\text{C.9})$$

Now, let

$$m_{q-} = \min \left(P [R_q^{c_q} = -1], P [R_q^{c'_q} = -1] \right)$$

and respectively

$$m_{q+} = \min \left(P [R_q^{c_q} = +1], P [R_q^{c'_q} = +1] \right)$$

Then, each of the above terms reduces to:

$$\left| \langle R_q^{c_q} \rangle - \langle R_q^{c'_q} \rangle \right| = 2 (1 - (m_{q+} + m_{q-})) \quad (\text{C.10})$$

Hence, following our previous corollary, the result follows. \square

C.2 Proof of proposition 1.46

What we want to prove is that for most empirical models we have:

$$\max_X \Delta_{C,C'}^*(X) = \sigma \quad (\text{C.11})$$

Definitions

First, we need to define all the terms in (C.11), and unify the notation used.

We start from a list of contexts \mathcal{C} (we want a rather definition of “context”, so a *context* will include the list of measurements + potential dependence to variables that depends on each individual context; that means that two identical lists of measurements can refer to two different contexts) from which we define the empirical model $e = (e^C)_{C \in \mathcal{C}}$

- Each of the e^C are probability distribution over possible outcomes in context C

- We will denote \mathcal{O}_C the set of possible outcomes in context C
- Given an observable X in the measurement context of C (write $X \in C$), we write $e_X^C = e^C|_X$ the marginal distribution corresponding to the observable X in the context C . Similarly, we define the set \mathcal{O}_X as the set of possible outcomes of the observable X .

A hidden variable model (HVM) of an empirical model e is here defined as $\Omega = (h = (h^\lambda)_{\lambda \in \Lambda}, p_\Lambda)$ where:

- Λ is the set of hidden/latent variables in the HVM
- For all hidden variable λ and context C , $h^{\lambda,C}$ is a probability distribution over \mathcal{O}_C ; therefore $(h^{\lambda,C})_{C \in \mathcal{C}}$ forms an empirical model.
- p_Λ is a probability distribution over Λ
- For every context C we have:

$$e^C = \sum_{\lambda \in \Lambda} p_\Lambda(\lambda) h^{\lambda,C} \quad (\text{C.12})$$

For all hidden variable λ in a HVM, we can decompose h^λ as:

$$h^\lambda = c_{NS}^\lambda h_{NS}^\lambda + (1 - c_{NS}^\lambda) h'^\lambda \quad (\text{C.13})$$

where h_{NS}^λ is no-signalling, and h'^λ can be any empirical model.

We then define the *signalling fraction* σ as:

$$\sigma = \min_{HVM} \max_{\lambda \in \Lambda} 1 - c_{NS}^\lambda \quad (\text{C.14})$$

In the M-contextuality framework (framework fundamentally related to the Cbd framework), we are interested in *canonical causal models* \mathcal{M} in which each of the individual observable X is associated with a random variable; the (different choices of) contexts are also modelled as a single random variable which can influence (all of the different) observable variables. In addition, we also define a latent variable Λ which is independent of the context variable but can also influence all of the observable variables. These models can themselves be viewed as hidden variable models

in the sense described above by setting:

$$h^{\lambda,C}(o) = Pr_{\mathcal{M}}[o \mid C, \lambda] \quad (\text{C.15})$$

and from where we can also recover (C.12). When obvious we will drop the \mathcal{M} subscript.

Without loss of generality, it is also enough to restrict ourselves to canonical models \mathcal{M} such that for all $\lambda \in \Lambda, C \in \mathcal{C}, X \in \mathcal{C}$ and $x \in \mathcal{O}_X$, we have $h_X^{\lambda,C}(x) \in \{0, 1\}$. As each of the $h_X^{\lambda,C}$ are probability distributions over \mathcal{O}_X , we can therefore define for each pair (λ, C) and observable $X \in \mathcal{C}$ a function $F_X : \Lambda \times \mathcal{C} \rightarrow \mathcal{O}_X$ such that:

$$F_X(\lambda, C) = x \iff h_X^{\lambda,C}(x) = 1 \quad (\text{C.16})$$

Given a canonical model \mathcal{M} , we can define the degree of direct influence from the (change of) context $C \leftrightarrow C'$ on the observable variable $X \in C \cup C'$ as:

$$\Delta_{C,C'}(X) = Pr[\{\lambda \mid F_X(\lambda, C) \neq F_X(\lambda, C')\}] \quad (\text{C.17})$$

We now introduce a couple of results from CbD and M-contextuality:

- For all observables X , we define:

$$Pr[e_X^C = e_X^{C'}] = \sum_{o \in \mathcal{O}_X} \min_{\tilde{C} \in \{C, C'\}} e_X^{\tilde{C}}(o) \quad (\text{C.18})$$

- The above equation can also be extended to the situation where more than two contexts intersect at the observable X as follows:

$$Pr[e_X^{C_1} = e_X^{C_2} = \dots = e_X^{C_n}] = \min_{(i,j) \in \{1,2,\dots,n\}^2, i \neq j} Pr[e_X^{C_i} = e_X^{C_j}] \quad (\text{C.19})$$

- For all canonical models \mathcal{M} , we always have:

$$\Delta_{C,C'}(X) \leq Pr[e_X^C \neq e_X^{C'}] = 1 - Pr[e_X^C = e_X^{C'}] \quad (\text{C.20})$$

- For any empirical model, for any observable X , there exist a canonical model such that:

$$\Delta_{C,C'}(X) = Pr[e_X^C \neq e_X^{C'}] \quad (\text{C.21})$$

i.e. the maximum in (C.20) can always be attained in a canonical model. We will also write:

$$\Delta_{C,C'}^*(X) = Pr \left[e_X^C \neq e_X^{C'} \right] = \max_{\mathcal{M}} \Delta_{C,C'}(X) \quad (\text{C.22})$$

Inequality 1 (the general case)

We first prove that for an empirical model:

$$\max_X \Delta_{C,C'}^*(X) \leq \sigma \quad (\text{C.23})$$

Proof. Suppose that there exists an observable X in an empirical model for which:

$$\Delta_{C,C'}^*(X) > \sigma = \max_{\lambda} 1 - c_{NS}^{\lambda} \quad (\text{C.24})$$

From M-contextuality, there exists a canonical model \mathcal{M} in which:

$$\Delta_{C,C'}(X) = \Delta_{C,C'}^*(X) = 1 - \sum_{x \in \mathcal{O}_X} \min_{\tilde{C} \in \{C,C'\}} e_X^{\tilde{C}}(x) \quad (\text{C.25})$$

Now, using the HVM corresponding to this canonical model, we have:

$$e_X^{\tilde{C}}(x) = \sum_{\lambda \in \Lambda} p_{\lambda}(\lambda) h_X^{\lambda, \tilde{C}}(x) \quad (\text{C.26})$$

so:

$$\Delta_{C,C'}(X) = 1 - \sum_{x \in \mathcal{O}_X} \min_{\tilde{C} \in \{C,C'\}} \sum_{\lambda \in \Lambda} p_{\lambda}(\lambda) h_X^{\lambda, \tilde{C}}(x) > \max_{\lambda} 1 - c_{NS}^{\lambda} = 1 - \min_{\lambda} c_{NS}^{\lambda} \quad (\text{C.27})$$

In turns that implies that:

$$\sum_{x \in \mathcal{O}_X} \min_{\tilde{C} \in \{C,C'\}} \sum_{\lambda \in \Lambda} p_{\lambda}(\lambda) h_X^{\lambda, \tilde{C}}(x) < \min_{\lambda} c_{NS}^{\lambda} \quad (\text{C.28})$$

We then decompose h^{λ} as the convex sum of a no-signalling and a signalling part:

$$\sum_{x \in \mathcal{O}_X} \min_{\tilde{C} \in \{C,C'\}} \sum_{\lambda \in \Lambda} p_{\lambda}(\lambda) \left[c_{NS}^{\lambda} h_{NS,X}^{\lambda, \tilde{C}}(x) + (1 - c_{NS}^{\lambda}) h_X^{\lambda, C}(x) \right] < \min_{\lambda} c_{NS}^{\lambda} \quad (\text{C.29})$$

Now, since $(1 - c_{NS}^\lambda) h_X^{\lambda,C}(x) > 0$, this implies that:

$$\sum_{x \in \mathcal{O}_X} \min_{\tilde{C} \in \{C, C'\}} \sum_{\lambda \in \Lambda} p_\lambda(\lambda) c_{NS}^\lambda h_{NS,X}^{\lambda, \tilde{C}}(x) < \min_{\lambda} c_{NS}^\lambda \quad (\text{C.30})$$

We then use the fact that h_{NS}^λ is no-signalling, and therefore:

$$h_{NS,X}^{\lambda,C}(x) = h_{NS,X}^{\lambda,C'}(x) \quad (\text{C.31})$$

So (C.30) simplifies to:

$$\sum_{x \in \mathcal{O}_X} \sum_{\lambda \in \Lambda} p_\lambda(\lambda) c_{NS}^\lambda h_{NS,X}^{\lambda,C}(x) = \sum_{\lambda \in \Lambda} p_\lambda(\lambda) c_{NS}^\lambda \sum_{x \in \mathcal{O}_X} h_{NS,X}^{\lambda,C}(x) < \min_{\lambda} c_{NS}^\lambda \quad (\text{C.32})$$

Now, $h_{NS,X}^{\lambda,C}$ is a probability distribution over \mathcal{O}_X , so $\sum_x h_{NS,X}^{\lambda,C}(x) = 1$, and:

$$\sum_{\lambda \in \Lambda} p_\lambda(\lambda) c_{NS}^\lambda < \min_{\lambda} c_{NS}^\lambda \quad (\text{C.33})$$

In addition, for all λ , we have $\min_{\lambda'} c_{NS}^\lambda \leq c_{NS}^\lambda$, so:

$$\sum_{\lambda \in \Lambda} p_\lambda(\lambda) \min_{\lambda'} c_{NS}^{\lambda'} \leq \sum_{\lambda \in \Lambda} p_\lambda(\lambda) c_{NS}^\lambda < \min_{\lambda} c_{NS}^\lambda \quad (\text{C.34})$$

and:

$$\sum_{\lambda \in \Lambda} p_\lambda(\lambda) \min_{\lambda'} c_{NS}^{\lambda'} = \min_{\lambda'} c_{NS}^{\lambda'} \sum_{\lambda \in \Lambda} p_\lambda(\lambda) = \min_{\lambda'} c_{NS}^{\lambda'} \quad (\text{C.35})$$

So we then obtain the contradiction:

$$\min_{\lambda'} c_{NS}^{\lambda'} < \min_{\lambda} c_{NS}^\lambda \quad (\text{C.36})$$

□

Inequality 2

The second inequality, i.e.:

$$\max_X \Delta_{C,C'}^*(X) \geq \sigma \quad (\text{C.37})$$

only holds in cases when the notion of consistent connectedness and no-signalling coincides.

Proof. We want to construct an HVM such that:

$$\max_{\lambda} c_{NS}^{\lambda} = \max_X \Delta_{C,C'}^*(X) \quad (\text{C.38})$$

We isolate the observable X such that $\max_{\tilde{X}} \Delta_{C,C'}^*(\tilde{X}) = \Delta_{C,C'}^*(X)$, and will denote for simplicity $\Delta = \Delta_{C,C'}^*(X)$. Then, from M-contextuality, we know that there exist a canonical model \mathcal{M} such that $\Delta_{C,C'}(X) = \Delta$. We will use this canonical model to create a HVM with a single hidden variable λ where:

$$h_{NS,Y}^{\lambda,D}(y) = h_{NS,Y}^{\lambda,D'}(y) = \frac{1}{1 - \Delta_{D,D'}(Y)} Pr_{\mathcal{M}} [\{\lambda \mid F_Y(\lambda, D) = F_Y(\lambda, D') = y\}] \quad (\text{C.39})$$

Note: if the observable Y is part of more than 2 contexts $D_1, D_2, D_3 \dots$, we will replace $\Delta_{D,D'}(Y)$ by $\min_{i,j} \Delta_{D_i,D_j}(Y)$, and $Pr_{\mathcal{M}} [\{\lambda \mid F_Y(\lambda, D) = F_Y(\lambda, D') = y\}]$ by:

$$Pr_{\mathcal{M}} [\{\lambda \mid F_Y(\lambda, D_1) = F_Y(\lambda, D_2) = F_Y(\lambda, D_3) = \dots = y\}]$$

Since we are only working with a single hidden variable, we will also drop the λ superscript.

We then show that this actually leads to a valid HVM.

Claim C.4. 1. $h_{NS,Y}$ is a probability distribution over \mathcal{O}_Y

2. For all $y \in \mathcal{O}_Y$:

$$(1 - \Delta) h_{NS,Y}(y) \leq e_Y^D(y), e_Y^{D'}(y) \quad (\text{C.40})$$

Proof. 1. By definition:

$$\Delta_{D,D'}(Y) = 1 - Pr_{\mathcal{M}} [\{\lambda \mid F_Y(\lambda, D) = F_Y(\lambda, D')\}] \quad (\text{C.41})$$

$$= 1 - \sum_{y \in \mathcal{O}_Y} Pr_{\mathcal{M}} [\{\lambda \mid F_Y(\lambda, D) = F_Y(\lambda, D') = y\}] \quad (\text{C.42})$$

Then

$$\sum_{y \in \mathcal{O}_Y} h_{NS,Y}(y) = \frac{1 - \sum_y Pr_{\mathcal{M}} [\{\lambda \mid F_Y(\lambda, D) = F_Y(\lambda, D') = y\}]}{1 - \sum_{y'} Pr_{\mathcal{M}} [\{\lambda \mid F_Y(\lambda, D) = F_Y(\lambda, D') = y'\}]} = 1 \quad (\text{C.43})$$

2. First recall that in the HVM $\left((g^\lambda)_{\lambda \in \Lambda}, q_\Lambda \right)$ we have:

$$e_Y^D(y) = \sum_{\lambda} q_\Lambda(\lambda) g_Y^{\lambda, D}(y) \quad (\text{C.44})$$

where $g_Y^{\lambda, D}(y) = 1$ iff $F_Y(\lambda, D) = y$ and $g_Y^{\lambda, D}(y) = 0$ otherwise. So, in fact:

$$e_Y^D(y) = Pr_{\mathcal{M}}[\{\lambda \mid F_Y(\lambda, D) = y\}] \quad (\text{C.45})$$

In addition, we have:

$$\Delta = \max_Y \max_{\mathcal{M}} \Delta_{D, D'}(Y) \quad (\text{C.46})$$

So:

$$\Delta \geq \Delta_{D, D'}(Y) \implies 1 - \Delta \leq 1 - \Delta_{D, D'}(Y) \quad (\text{C.47})$$

$$\implies \frac{1 - \Delta}{1 - \Delta_{D, D'}(Y)} \leq 1 \quad (\text{C.48})$$

Now:

$$(1 - \Delta)h_{NS, Y}(y) = \frac{1 - \Delta}{1 - \Delta_{D, D'}(Y)} Pr_{\mathcal{M}}[\{\lambda \mid F_Y(\lambda, D) = F_Y(\lambda, D') = y\}] \quad (\text{C.49})$$

$$\leq Pr_{\mathcal{M}}[\{\lambda \mid F_Y(\lambda, D) = F_Y(\lambda, D') = y\}] \quad (\text{C.50})$$

$$\leq Pr_{\mathcal{M}}[\{\lambda \mid F_Y(\lambda, D) = y\}] = e_Y^D(y) \quad (\text{C.51})$$

since $F_Y(\lambda, D) = F_Y(\lambda, D') = y \implies F_Y(\lambda, D) = y$.

A similar proof can be done for D' . □

Claim C.5. We can define $h_{NS, Y}^D, h_{NS, Y}^{D'}$ distributions over \mathcal{O}_D and $\mathcal{O}_{D'}$ respectively such that:

$$1. h_{NS, Y}^D(o|_Y) = h_{NS, Y}^{D'}(o|_Y) = h_{NS, Y}(o|_Y)$$

2. For all $o \in \mathcal{O}_D$:

$$(1 - \Delta)h_{NS}^D(o) \leq e^D(o) \quad (\text{C.52})$$

(and similarly for D').

Proof. We'll only show this for the context D (but the same applies for D').

We define:

$$h_{NS}^D(o) = \frac{e^D(o)}{e_Y^D(o|Y)} h_{NS,Y}(o|Y) \quad (\text{C.53})$$

This defines a probability distribution since:

$$\sum_{o \in \mathcal{O}_D} h_{NS}^D(o) = \sum_{o \in \mathcal{O}_D} \frac{e^D(o)}{e_Y^D(o|Y)} h_{NS,Y}(o|Y) \quad (\text{C.54})$$

$$= \sum_{y \in \mathcal{O}_Y} \sum_{o \in \mathcal{O}_D | o|Y=y} \frac{e^D(o)}{e_Y^D(y)} h_{NS,Y}(y) \quad (\text{C.55})$$

$$= \sum_{y \in \mathcal{O}_Y} \frac{h_{NS,Y}(y)}{e_Y^D(y)} \sum_{o \in \mathcal{O}_D | o|Y=y} e^D(o) = \sum_{y \in \mathcal{O}_Y} \frac{h_{NS,Y}(y)}{e_Y^D(y)} e^D(y) \quad (\text{C.56})$$

$$= \sum_{y \in \mathcal{O}_Y} h_{NS,Y}(y) = 1 \quad (\text{C.57})$$

1. We can just check that:

$$h_{NS|Y}^D(y) = \sum_{o \in \mathcal{O}_D | o|Y=y} \frac{e^D(o)}{e_Y^D(y)} h_{NS,Y}(y) = h_{NS,Y}(y) \quad (\text{C.58})$$

2. Similarly, we just check that:

$$(1 - \Delta) h_{NS,Y}^D(o) = \frac{1 - \Delta}{1 - \Delta_{D,D'}(Y)} \frac{e^D(o)}{e_Y^D(o|Y)} \Pr_{\mathcal{M}}[\{\lambda \mid F_Y(\lambda, D) = F_Y(\lambda, D') = y\}] \quad (\text{C.59})$$

$$\leq e^D(o) \frac{\Pr_{\mathcal{M}}[\{\lambda \mid F_Y(\lambda, D) = F_Y(\lambda, D') = y\}]}{e_Y^D(o|Y)} \quad (\text{C.60})$$

$$\leq e^D(o) \quad (\text{C.61})$$

□

Corollary C.6. *There is some empirical model h' such that:*

$$e = (1 - \Delta) h_{NS} + \Delta h' \quad (\text{C.62})$$

Proof. We have already shown that $(1 - \Delta) h_{NS} \leq e$. We then define:

$$h' = \frac{1}{\Delta} (e - (1 - \Delta) h_{NS}) \geq 0 \quad (\text{C.63})$$

This is an empirical mode since for all contexts D , we have:

$$\sum_{o \in \mathcal{O}_D} h^{ID}(o) = \frac{1}{\Delta} \left[\sum_{o \in \mathcal{O}_D} e^D(o) - (1 - \Delta) \sum_{o \in \mathcal{O}_D} h_{NS}^D(o) \right] = \frac{1}{\Delta} [1 - (1 - \Delta)] = 1 \quad (\text{C.64})$$

□

□

Why it doesn't work when $|C \cap C'| > 1$

The proof of (C.37) relies on the fact that there is for each observable X a function $F_X(\lambda, C) : \Lambda \times \mathcal{C} \rightarrow X$ which defines the probability distribution of the hidden-variable distributions h^λ . Now, if we were in a situation when $|C \cap C'| \geq 2$, then we would not only (in the sheaf-theoretic framework) check the no-signalling condition on each of the $X \in C \cap C'$ but also for each $S \subseteq C \cap C'$.

The above reasoning could easily be extended if there was a function $F_S : \Lambda \times \mathcal{C} \rightarrow \prod_{X \in S} \mathcal{O}_X$, which restrict to F_X by post-composing with a projection operator (recall that we want *all* restrictions to be well-defined). But then, we would also have:

$$1 - \Delta_{C,C'}(\{X, Y\}) = Pr [\{\lambda | F_{\{X,Y\}}(\lambda, C) = F_{\{X,Y\}}(\lambda, C')\}] \quad (\text{C.65})$$

$$\begin{aligned} &\leq Pr [\{\lambda | \pi_X \circ F_{\{X,Y\}}(\lambda, C) = \pi_X \circ F_{\{X,Y\}}(\lambda, C')\}] \\ &= \Delta_{C,C'}(X) \end{aligned} \quad (\text{C.66})$$

And therefore, if \mathcal{X} denotes the set of observables:

$$\min_{S \in \mathcal{P}(\mathcal{X})} \Delta_{C,C'}(S) \geq \min_{X \in \mathcal{X}} \Delta_{C,C'}(X) \quad (\text{C.67})$$

in every canonical model. Hence, there could exist a model such that:

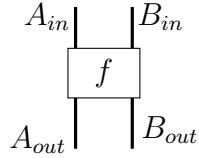
$$\max_{\mathcal{M}} \max_{S \subseteq \mathcal{X}} \Delta_{C,C'}(S) \geq \max_{\lambda} 1 - c_{NS}^\lambda > \max_{\mathcal{M}} \max_{X \in \mathcal{X}} \Delta_{C,C'}(X) \quad (\text{C.68})$$

C.3 Equivalence between causality notions

We start by discussing the equivalence between distributions of causal *functions* and causality as defined in terms of causal *processes* for bipartite systems, and we

will (without loss of generality) consider the causal order $A \preceq B$. We will then extend this to arbitrary causal orders.

First, we need to state that distributions of causal functions in $\mathcal{D}_{\mathbb{R}_+}(\mathcal{L}_{A \preceq B}(\{A, B\}))$ live in classical probability spaces while the output of quantum processes such as the ones in (1.111) or (1.113) are quantum states. Hence, we will here show an equivalence between causality as described in Section 1.2 and the *measurement statistics* of causal processes. Namely, for a quantum circuit:



and the interpretation of input a, b and outputs o_a, o_b as quantum states $|a\rangle, |b\rangle, |o_a\rangle, |o_b\rangle$, we define the conditional probability:

$$P[o_a, o_b | a, b] = \begin{array}{c} \triangle a \quad \triangle b \\ | \quad | \\ A_{in} \quad B_{in} \\ \boxed{f} \\ | \quad | \\ A_{out} \quad B_{out} \\ \triangle o_a \quad \triangle o_b \end{array} \quad (C.69)$$

Moreover, assuming that the set $\{|o_B\rangle\}$ spans the entire space B_{out} , i.e. that:

$$\sum_{o_b} \begin{array}{c} | \\ \triangle o_b \end{array} = \underline{\underline{1}} \quad (C.70)$$

This is reasonable as we expect that, at each run, some outcome is detected. Furthermore, it should also be the case that:

$$\begin{array}{c} \triangle b \\ | \\ \underline{\underline{1}} \end{array} = 1 \quad (C.71)$$

i.e. discarding a state is deterministic.

Using the above assumptions, the operational condition of equation (1.111) for the causal order $A \preceq B$ implies that:

The diagrammatic equation (C.72) consists of three parts connected by equals signs. The first part shows a box labeled f with two input wires labeled A_{in} and B_{in} entering from the top. The A_{in} wire is connected to a triangle labeled a , and the B_{in} wire is connected to a triangle labeled b . Two output wires labeled A_{out} and B_{out} exit from the bottom. The A_{out} wire is connected to a triangle labeled o_a , and the B_{out} wire is connected to a ground symbol. The second part shows the same box f with the same inputs, but the B_{out} wire is connected to a triangle labeled o_b instead of ground. This part is preceded by a summation symbol \sum_{o_b} . The third part shows a box labeled \tilde{f} with the same inputs A_{in} and B_{in} . The A_{out} wire is connected to a triangle labeled o_a , and the B_{out} wire is connected to ground. This part is preceded by an equals sign. The entire equation is labeled (C.72) on the right.

which, in terms of conditional probabilities translate as:

$$\sum_{o_b} P[o_a, o_b | a, b] = P[o_a | a] \quad (\text{C.73})$$

Now, let F being a set of functions $f : A \times B \rightarrow O_A \times O_B$ corresponding to events then, for any section $\mu \in \mathcal{D}_{\mathbb{R}_+}(F)$, we have:

$$P[o_a, o_b | a, b] = \sum_{f \text{ s.t. } f(a,b)=(o_a,o_b)} \mu(f) \quad (\text{C.74})$$

and:

$$P[o_a | a] = \sum_{f \text{ s.t. } f|_A(a)=(o_a)} \mu(f) \quad (\text{C.75})$$

Therefore, we have μ compatible with the causal order $A \preceq B$ with respect to the operational definition of equation (1.111) iff:

$$\sum_{o_b} \sum_{f \text{ s.t. } f(a,b)=(o_a,o_b)} \mu(f) = \sum_{f \text{ s.t. } f|_A(a)=(o_a)} \mu(f) \quad (\text{C.76})$$

Now, since the LHS of equation (C.76) is marginalising o_b , the free variables on the LHS are a, b, o_a , whereas the free variables in the RHS are a, o_a . Therefore, the above equation is satisfied iff the values of $f|_A$ do not depend on the value of b , which is exactly the condition for causality of functions with respect to the causal order $A \preceq B$.

In order to extend this result to arbitrary causal order, we employ a trick from [104], namely to split any given causal order (Σ, \preceq) as a coarse-grained causal order

$A \preceq B$, where A consists on the parties $\Sigma\{B\}$, and B is a *maximal box*, i.e. a party which does not influence any other one. Then, we know from the above result that a circuit is compatible with the causal order $A \preceq B$ iff the statistics of the measurements correspond to a section $\mu \in \mathcal{D}_{\mathbb{R}_+}(F)$, where F is a set of causal functions with respect to $A \preceq B$. Therefore, we need to check that $\mu|_A$ is still compatible with the causal order Σ . We then proceed to isolate a new $B' \in A$ such that B' is maximal in the reduced causal order $\Sigma\{B\}$. We then keep going until the reduced scenario only consist of 2 parties.

C.4 Proof of proposition 3.12

In order to prove the formula for the causal fraction in (2,2,2)-Bell scenarios, we start by proving a more general equation that the causal fraction needs to satisfy, in any empirical model.

Proposition C.7. *For a family of probability distributions where the causal order is not known, an upper bound of the causal fraction can be calculated as follows^[1]:*

$$\gamma \leq \min_{U,V} 1 - |e_{\underline{i}}|_{U|_{U \cap V}}(\underline{a}) - e_{\underline{i}}|_{V|_{U \cap V}}(\underline{a})| \quad (\text{C.77})$$

where $e_{\underline{i}}|_{U|_{U \cap V}}$ corresponds to the restriction of $e_{\underline{i}}$ to first U and then from U to $U \cap V$ (and similarly for $e_{\underline{i}}|_{V|_{U \cap V}}$).

Proof. For every causal empirical causal model e^Ω with respect to a causal scenario $\Sigma = (\Omega, \underline{I}, \underline{Q})$, if we have $\gamma \cdot e^\Omega \preceq e$, then both:

$$\gamma \cdot e_{\underline{i}}^\Omega|_{U \cap V}(\underline{a}) \leq e_{\underline{i}}|_{U|_{U \cap V}}(\underline{a}) \quad (\text{C.78})$$

and

$$\gamma \cdot e_{\underline{i}}^\Omega|_{U \cap V}(\underline{a}) \leq e_{\underline{i}}|_{V|_{U \cap V}}(\underline{a}) \quad (\text{C.79})$$

So:

$$\gamma \cdot e_{\underline{i}}^\Omega|_{U \cap V}(\underline{a}) \leq \min_{X \in \{U,V\}} e_{\underline{i}}|_X|_{U \cap V}(\underline{a}) \quad (\text{C.80})$$

^[1]Note: the order of the restrictions is from left to right.

Now, since e_i are probability distributions:

$$1 - e_i|_X|_{U \cap V}(\varrho) = \sum_{\varrho' \neq \varrho} e_i|_X|_{U \cap V}(\varrho') \quad (\text{C.81})$$

and similarly for e^Ω . Therefore, using $\gamma e^\Omega \preceq e$ once again:

$$\gamma \left(1 - e_i^\Omega|_X|_{U \cap V}(\varrho) \right) \leq \min_{X \in \{U, V\}} 1 - e_i|_X|_{U \cap V}(\varrho) = 1 - \max_{X \in \{U, V\}} e_i|_X|_{U \cap V}(\varrho) \quad (\text{C.82})$$

Then, writing $m_- = \min_{X \in \{U, V\}} e_i|_X|_{U \cap V}(\varrho)$ and $m_+ = \max_{X \in \{U, V\}} e_i|_X|_{U \cap V}(\varrho)$ for simplicity, we use (C.80) and (C.82) to get:

$$\gamma \leq 1 - m_+ + m_- \quad (\text{C.83})$$

Now, using binary minima and maxima this reduces to:

$$\gamma \leq 1 - |e_i|_U|_{U \cap V}(\varrho) - e_i|_V|_{U \cap V}(\varrho)| \quad (\text{C.84})$$

And since this has to be the case for all $U, V \in \mathcal{L}$, the claimed inequality has to hold. \square

Let's now describe a construction of a causal empirical model e^Ω which satisfies $\gamma \cdot e^\Omega \preceq e$, for any given (2,2,2) Bell-type model e , where γ is given as in (3.9). This would therefore give proof that the above inequality can be saturated, and therefore that the causal fraction of such models can be known with certainty.

We start by constructing a probability distribution for the event A as follows. For any $a \in I_A$, we select $o_A^* \in O$ such that:

$$\min_{b \in I_B} e_{(a,b)}|_A(o_A^*) = \min_{o \in O} \min_{b \in I_B} e_{(a,b)}|_A(o) \quad (\text{C.85})$$

and set:

$$e_{(a,b)}^\Omega|_A(o_A^*) = \frac{\min_{b \in I_B} e_{(a,b)}|_A(o_A^*)}{\gamma} \quad (\text{C.86})$$

and $e_{(a,b)}^\Omega|_A(\neg o_A^*) = 1 - e_{(a,b)}^\Omega|_A(o_A^*)$. Then we have:

$$\gamma \cdot e_{(a,b)}^\Omega|_A(o) \leq e_{(a,b)}|_A(o) \quad (\text{C.87})$$

for all $(a, b) \in I_A \times I_B$, and for all possible outcome $o \in O$.

One can then extend this distribution to the lower set $A \rightarrow B = \Omega$ by setting, for example:

$$e_{(a,b)}^\Omega(o_A, o_B) = \frac{e_{(a,b)}^\Omega(o_A, o_B)}{e_{(a,b)}|_A(o_A)} e_{(a,b)}^\Omega|_A(o_A) \quad (\text{C.88})$$

It is routine to check that this construction leads to a valid empirical model e^Ω , which does indeed satisfy $\gamma \cdot e^\Omega \preceq e$.

C.5 Proof of proposition 4.4

Without loss of generality, let us assume that:

$$\gamma = 1 - e_{(a_1, b_1)}|_A(0) + e_{(a_1, b_2)}|_A(0) \quad (\text{C.89})$$

(If this is not the case, it is possible to form a new empirical model by relabelling inputs of A and B such that the above equation is valid). We then write for simplicity:

$$\alpha_{2i_A+i_B-3} = e_{(a_{i_A}, b_{i_B})}|_A(0) \quad (\text{C.90})$$

Let us also write:

$$e_{(a_{i_A}, b_{i_B})}(o_A, o_B) = \alpha_{2i_A+i_B-3, 2o_A+o_B} \quad (\text{C.91})$$

We are then looking for a target empirical e_Σ of the form:

	(0,0)	(0,1)	(1,0)	(1,1)
a_1, b_1	$p_1 x_1$	$(1 - p_1)x_1$	$p_2(1 - x_1)$	$(1 - p_2)(1 - x_1)$
a_1, b_2	$q_1 x_1$	$(1 - q_1)x_1$	$q_2(1 - x_1)$	$(1 - q_2)(1 - x_1)$
a_2, b_1	$r_1 x_2$	$(1 - r_1)x_2$	$r_2(1 - x_2)$	$(1 - r_2)(1 - x_2)$
a_2, b_2	$s_1 x_2$	$(1 - s_1)x_2$	$s_2(1 - x_2)$	$(1 - s_2)(1 - x_2)$

such that:

$$\gamma \cdot e_\Sigma \leq e \quad (\text{C.92})$$

Some intermediate results about the target empirical model e_Σ are:

1. $x_1 = \frac{\alpha_{0,3} - (1-\gamma)}{\gamma} = \frac{\alpha_1}{\gamma}$
2. $p_2 = 1 - \frac{\alpha_{0,3}}{\gamma(1-x_1)} = \frac{\alpha_{0,2}}{\gamma(1-x_1)}$

$$3. q_1 = 1 - \frac{\alpha_{1,1}}{\gamma x_1} = \frac{\alpha_{1,0}}{\gamma x_1}$$

Proof. 1. Since $\gamma = 1 - \alpha_0 + \alpha_1$, and $\gamma \geq 0$ then we have:

$$\alpha_0 \geq \alpha_1 \tag{C.93}$$

Now, from (C.92), we have have both of:

$$\gamma \cdot x_1 \leq \alpha_1 \iff x_1 \leq \frac{\alpha_1}{\gamma} \tag{C.94}$$

$$\gamma \cdot (1 - x_1) \leq 1 - \alpha_0 \iff \frac{\gamma - 1 + \alpha_0}{\gamma} \leq x_1 \tag{C.95}$$

Using (C.92) again, we have $\alpha_1 = \gamma - 1 + \alpha_0$ so in fact:

$$x_1 = \frac{\alpha_1}{\gamma} = \frac{\alpha_0 - (1 - \gamma)}{\gamma} \tag{C.96}$$

2. From (C.92), we know that:

$$\gamma \cdot p_2(1 - x_1) \leq \alpha_{0,2} \tag{C.97}$$

$$\gamma \cdot (1 - p_2)(1 - x_1) \leq \alpha_{0,3} \tag{C.98}$$

$$\tag{C.99}$$

iff:

$$1 - \frac{\alpha_{0,3}}{\gamma(1 - x_1)} \leq p_2 \leq \frac{\alpha_{0,2}}{\gamma(1 - x_1)} \tag{C.100}$$

Now, if (C.96) holds, then we have:

$$\gamma \cdot (1 - x_1) = 1 - \alpha_0 \tag{C.101}$$

and from the definition of α_0 , we get:

$$1 - \alpha_0 = \alpha_{0,2} + \alpha_{0,3} \tag{C.102}$$

So:

$$1 - \frac{\alpha_{0,3}}{\gamma(1 - x_1)} = \frac{\alpha_{0,2}}{1 - \alpha_0} \tag{C.103}$$

and therefore:

$$p_2 = \frac{\alpha_{0,2}}{\gamma(1-x_1)} = 1 - \frac{\alpha_{0,3}}{\gamma(1-x_1)} \quad (\text{C.104})$$

3. Similarly, given (C.92), we have:

$$\gamma q_1 x_1 \leq \alpha_{1,0} \quad (\text{C.105})$$

$$\gamma(1-q_1)x_1 \leq \alpha_{1,1} \quad (\text{C.106})$$

iff:

$$1 - \frac{\alpha_{1,1}}{\gamma x_1} \leq q_1 \leq \frac{\alpha_{1,0}}{\gamma} \quad (\text{C.107})$$

and using (C.96), we have:

$$\gamma x_1 = \alpha_1 \quad (\text{C.108})$$

and from the definition of α_1 :

$$\alpha_1 = \alpha_{1,0} + \alpha_{1,1} \quad (\text{C.109})$$

So:

$$1 - \frac{\alpha_{1,1}}{\gamma x_1} = \frac{\alpha_{1,0}}{\gamma x_1} \quad (\text{C.110})$$

Therefore:

$$q_1 = 1 - \frac{\alpha_{1,1}}{\gamma x_1} = \frac{\alpha_{1,0}}{\gamma x_1} \quad (\text{C.111})$$

□

We then write:

$$V_{a_k, b_j} = \frac{1}{2} \sum_o |e_{\Sigma, (a_k, b_j)}(o) - e_{(a_k, b_j)}(o)| \quad (\text{C.112})$$

so:

$$\min_{e_\Sigma} TV(e_\Sigma, e) = \min_{(p_i, q_i, r_i, s_i)_{i \in \{1,2\}}} \max_{(k,l) \in \{1,2\}^2} V_{a_k, b_l} \quad (\text{C.113})$$

Then, using (C.104) and (C.111), we get:

$$V_{a_1, b_1} = \frac{1}{2} \left[(1 - \alpha_0) \frac{1 - \gamma}{\gamma} + f(p_1) \right] \quad (\text{C.114})$$

$$V_{a_1, b_2} = \frac{1}{2} \left[\alpha_1 \frac{1 - \gamma}{\gamma} + g(q_2) \right] \quad (\text{C.115})$$

where:

$$f(p_1) = \begin{cases} \alpha_{0,0} - \alpha_{0,1} - 2p_1x_1 + x_1 & \text{if } p_1 \in \left[1 - \frac{\alpha_{0,1}}{\gamma x_1}, 1 - \frac{\alpha_{0,1}}{x_1}\right] \\ \alpha_0 - x_1 & \text{if } p_1 \in \left[1 - \frac{\alpha_{0,1}}{x_1}, \frac{\alpha_{0,0}}{x_1}\right] \\ \alpha_{0,1} - \alpha_{0,0} + 2p_1x_1 - x_1 & \text{if } p_1 \in \left[\frac{\alpha_{0,0}}{x_1}, \frac{\alpha_{0,0}}{\gamma x_1}\right] \end{cases} \quad (\text{C.116})$$

and:

$$g(q_2) = \begin{cases} \alpha_{1,2} - \alpha_{1,3} - 2(1-x_1)q_2 + 1 - x_1 & \text{if } q_2 \in \left[1 - \frac{\alpha_{1,3}}{\gamma(1-x_1)}, 1 - \frac{\alpha_{1,3}}{1-x_1}\right] \\ x_1 - \alpha_1 & \text{if } q_2 \in \left[1 - \frac{\alpha_{1,3}}{1-x_1}, \frac{\alpha_{1,2}}{1-x_1}\right] \\ \alpha_{1,3} - \alpha_{1,2} + 2(1-x_1)q_2 - 1 + x_1 & \text{if } q_2 \in \left[\frac{\alpha_{1,2}}{1-x_1}, \frac{\alpha_{1,2}}{\gamma(1-x_1)}\right] \end{cases} \quad (\text{C.117})$$

which are both continuous functions with minima:

$$\min_{p_1} f(p_1) = \alpha_0 - x_1 = \frac{1-\gamma}{\gamma} (1 - \alpha_0) \quad (\text{C.118})$$

$$\min_{q_2} g(q_2) = x_1 - \alpha_1 = \frac{1-\gamma}{\gamma} \alpha_1 \quad (\text{C.119})$$

For a_2, b_1 , we consider the case where:

1. $r_1x_2 = \alpha_{2,0}$
2. $(1-r_1)x_2 = \alpha_{2,1}$
3. $r_1(1-x_2) = \alpha_{2,2}$
4. $(1-r_1)(1-x_2) = \alpha_{2,3}$

Therefore:

$$V_{a_2, b_1} = 0 \quad (\text{C.120})$$

Note that this is possible since this would imply that:

$$x_2 = \alpha_2 \quad (\text{C.121})$$

and:

$$\gamma\alpha_2 \leq \min_{k \in \{2,3\}} \alpha_k \quad (\text{C.122})$$

from (C.92) and similarly:

$$\gamma(1 - \alpha_2) \leq 1 - \max_{k \in \{2,3\}} \alpha_k \quad (\text{C.123})$$

And for r_1 , we have:

$$\frac{\alpha_{2,0}}{\alpha_2} \leq \frac{\alpha_{2,0}}{\gamma\alpha_2} \quad (\text{C.124})$$

and:

$$\frac{\alpha_{2,0}}{\alpha_2} = 1 - \frac{1 - \alpha_{2,1}}{\alpha_2} \geq 1 - \frac{1 - \alpha_{2,1}}{\gamma\alpha_2} \quad (\text{C.125})$$

(and similarly for r_2).

Now, for a_2, b_2 , we have:

$$\alpha_2 = x_2 \geq \alpha_3 \iff \frac{\alpha_2 - \alpha_{2,1}}{x_2} \geq \frac{\alpha_{2,0}}{x_2} \quad (\text{C.126})$$

so $\left[1 - \frac{\alpha_{2,1}}{\alpha_2}, \frac{\alpha_{2,0}}{\alpha_2}\right]$ is a non-empty interval iff $\alpha_2 \leq \alpha_3$ and in which case:

$$s_1 \in \left[1 - \frac{\alpha_{2,1}}{\alpha_2}, \frac{\alpha_{2,0}}{\alpha_2}\right] \implies s_1 x_2 \geq \alpha_{2,0} \wedge (1 - s_1)x_2 \geq \alpha_{2,1} \quad (\text{C.127})$$

So in this case:

$$V_{a_2, b_2} = \frac{1}{2} [\alpha_2 - \alpha_3 + |s_2(1 - \alpha_2) - \alpha_{2,2}| + |(1 - s_2)(1 - \alpha_2) - \alpha_{2,3}] \quad (\text{C.128})$$

$$= \frac{1}{2} [\alpha_2 - \alpha_3 + h(s_2)] \quad (\text{C.129})$$

where:

$$h(s_2) = \begin{cases} (1 - 2s_2)(1 - \alpha_2) + \alpha_{2,3} - \alpha_{2,2} & \text{if } s_2 \in \left[1 - \frac{\alpha_{2,3}}{\gamma(1-x_2)}, 1 - \frac{\alpha_{2,3}}{1-x_2}\right] \\ \alpha_2 - \alpha_3 & \text{if } s_2 \in \left[1 - \frac{\alpha_{2,3}}{1-x_2}, \frac{\alpha_{2,2}}{1-x_2}\right] \\ (2s_2 - 1)(1 - \alpha_2) + \alpha_{2,2} - \alpha_{2,3} & \text{if } s_2 \in \left[\frac{\alpha_{2,2}}{1-x_2}, \frac{\alpha_{2,2}}{\gamma(1-x_2)}\right] \end{cases} \quad (\text{C.130})$$

Now, since:

$$\gamma = \min \{1 - \alpha_0 + \alpha_1, 1 - \alpha_2 + \alpha_3\} \quad (\text{C.131})$$

then:

$$\alpha_2 - \alpha_3 \leq \frac{1 - \gamma}{\gamma} \min \{1 - \alpha_0, \alpha_1\} \quad (\text{C.132})$$

Hence:

$$\max_{(k,l) \in \{1,2\}^2} V_{a_k, b_l} = \frac{1-\gamma}{\gamma} \min\{1-\alpha_0, \alpha_1\} \quad (\text{C.133})$$

(and similar reasoning can be made in the case $\alpha_3 \leq \alpha_2$).

Thus, we then obtain the bound:

$$\min_{e_\Sigma} TV(e_\Sigma, e) \leq \frac{1-\gamma}{\gamma} \min\{1-\alpha_0, \alpha_1\} \quad (\text{C.134})$$

C.6 Proof of proposition 5.10

Each of our contexts includes exactly one less word as the next one. As a result given a pair of successive contexts, we can without loss of generality consider the 2-context scenario as an $\{m, mw\}$ scenario. The no-signalling condition of the model is then as follows:

$$e_{mw|_m} = e_m \quad (\text{C.135})$$

Given an arbitrary 2-context empirical model as above, we want to find the following decomposition:

$$e = \text{NSF} \cdot e_{NS} + \text{SF} \cdot e' \quad (\text{C.136})$$

where e_{NS} is the maximum possible across all such decompositions. Let us assume, as above, that m is a single ‘‘observable’’ with possible outcomes in $O^m = \{0, \dots, n^{|m|} - 1\}$, where $|m|$ is the number of words in the context m . Our first goal is to find a distribution for m in e_{NS} , which satisfies the following for all $o_i \in O^m$:

$$\text{NSF} e_{NS, m}(o_i) \leq \min(e_{mw|_m}(o_i), e_m(o_i)) \quad (\text{C.137})$$

From the above it follows that

$$\sum_{o_i} \text{NSF} e_{NS, m}(o_i) = \text{NSF} \leq \sum_{o_i} \min(e_{mw|_m}(o_i), e_m(o_i)) \quad (\text{C.138})$$

One can always construct an empirical model e_{NS} such that:

$$\sum_{o_i} \min(e_{mw|_m}(o_i), e_m(o_i)) e_{NS} \leq e \quad (\text{C.139})$$

In order to see this, first observe that the probability distribution of $e_{NS, m}$ can be

constructed by first relabeling the outcomes to o_{i_k} for $0 \leq k \leq n^{|m|} - 1$ such that for $N = n^{|m|} - 1$, the following holds:

$$\begin{aligned} \min(e_{mw|_m}(o_{i_0}), e_m(o_{i_0})) &\leq \min(e_{mw|_m}(o_{i_1}), e_m(o_{i_1})) \\ &\leq \dots \\ &\leq \min(e_{mw|_m}(o_{i_N}), e_m(o_{i_N})) \end{aligned} \quad (\text{C.140})$$

Then, we take σ to be $\sum_{o_i} \min(e_{mw|_m}(o_i), e_m(o_i))$ and set:

$$e_{NS,m}(o_{i_0}) = \frac{\min(e_{mw|_m}(o_{i_0}), e_m(o_{i_0}))}{\sigma} \quad (\text{C.141})$$

We can then inductively define:

$$e_{NS,m}(o_{i_k}) = \min\left(\frac{\min(e_{mw|_m}(o_{i_0}), e_m(o_{i_0}))}{\sigma}, 1 - \sum_{j=0}^{k-1} e_{NS,m}(o_{i_j})\right) \quad (\text{C.142})$$

From this definition, we know that for all k , we have the following:

$$\sigma e_{NS,m}(o_{i_k}) \leq e_{mw|_m}(o_{i_k}), e_m(o_{i_k}) \quad (\text{C.143})$$

In addition, the above forms a valid probability distribution as, if there exists a k such that $e_{NS,m}(o_{i_k}) = 1 - \sum_{j=0}^{k-1} e_{NS,m}(o_{i_j})$, then $e_{NS,m}(o_{i_{k'}}) = 0$ for all $k' > k$ and therefore:

$$\sum_k e_{NS,m}(o_{i_k}) = 1 \quad (\text{C.144})$$

Similarly, if for all k , $e_{NS,m}(o_{i_k}) = \frac{\min(e_{mw|_m}(o_{i_0}), e_m(o_{i_0}))}{\sigma}$, then by definition of σ , we also have (C.144). We extend this probability distribution to an empirical model over $\{m, mw\}$, by defining:

$$e_{NS,mw}(o_m, o_w) = e_{mw}(o_m, o_w) \frac{e_{NS,m}(o_m)}{e_{mw|_m}(o_m)} \quad (\text{C.145})$$

It is now easy to show that $e_{NS,mw}|_m = e_{NS,m}$ and in addition, we have:

$$\sigma e_{NS,mw}(o_m, o_w) = e_{mw}(o_m, o_w) \frac{\sigma e_{NS,m}(o_m)}{e_{mw|_m}(o_m)} \leq e_{mw}(o_m, o_w) \quad (\text{C.146})$$

As a result of the above calculations, the signalling fraction can be computed as follows:

$$\text{SF} = 1 - \sum_o \min(e_{mw|_m}(o), e_m(o)) \quad (\text{C.147})$$

Chapter D

LEXICAL AMBIGUITY DATASET

D.1 List of ambiguous words

Ambiguous nouns

Noun	Ambiguity	Definitions	Examples
Atmosphere	Polysemous	Layer of gas around the Earth/a planet (literal only) Feeling/Mood of a place or situation (figurative only)	<i>pollution of the atmosphere</i> <i>a relaxing atmosphere</i>
Band	Homonymous	Group of musicians playing together A thin flat piece of material or a range of values	<i>a rock or blues band</i> <i>an elastic band, the 18-25 age band</i>
Bank	Homonymous	A financial institution Sloping raised land, especially along the sides of a river	<i>bank accounts</i> <i>the bank of a river</i>
Bark	Homonymous	The hard outer covering of a tree Loud, rough noise (usually that dogs and animals make)	<i>tree bark, red bark</i> <i>a loud bark</i>

Noun	Ambiguity	Definitions	Examples
Beam	Homonymous	Line of light/radiation that shines from an object Long, thick piece of wood, metal, or concrete, especially used to support weight in a building or other structure	<i>beam of a torch, an electron beam</i> <i>wooden beams</i>
Book	Polysemous	Set of pages fastened together inside a cover (physical object) (Usually the content of) a written text that can be published in printed or electronic form	<i>a hardback book, a heavy book</i> <i>an interesting book, reading a book</i>
Boxer	Homonymous	An athlete practising boxing A specific breed of dog	<i>a heavyweight boxer</i> <i>the boxer is a hunting mastiff</i>
Cabinet	Homonymous	A small group of the most important people in government A piece of furniture with shelves, cupboards, or drawers	<i>a cabinet minister, a cabinet reshuffle</i> <i>a filing cabinet, a glass-fronted cabinet</i>
Chicken	Polysemous	Type of bird, usually kept in farms The meat of the bird of the same name	<i>a male/female chicken</i> <i>roast chicken, fried chicken</i>
Coach	Homonymous	Someone whose job is to teach people A motor vehicle like a bus	<i>a sport coach</i> <i>a coach trip</i>
Coat	Polysemous	Piece of clothing, usually used for warmth A layer of a substance	<i>a fur coat</i> <i>a coat of paint</i>
Cotton	Polysemous	Plants that produces flowers from which fabric can be made Material used in textile or cosmetic industry	<i>cotton fields, cotton plants</i> <i>a cotton shirt, cotton pads</i>

Noun	Ambiguity	Definitions	Examples
Fall	Homonymous	The fact of the size, amount, or strength of something getting lower In American English, the season between summer and winter (in British English: autumn)	<i>the fall in prices, the rise and fall</i> <i>fall colours, a fall day</i>
Film	Polysemous	A movie A thin layer of material (physical object)	<i>watching a film, a horror film</i> <i>a plastic film, a film of oil</i>
Glasses	Polysemous	Container for drinks Devices which are used to improve eyesight	<i>a wine glass, a glass of water</i> <i>a pair of glasses, reading glasses</i>
Iron	Polysemous	Metal that rusts / chemical element Device used to flatten / smooth clothes	<i>an iron chain, an iron deficiency</i> <i>a steam iron, a travel iron</i>
Letter	Polysemous	A written message sent to someone Symbol used in written language	<i>a love letter</i> <i>the letter F</i>
Library	Polysemous	A building that has a collections of books for people to borrow A collection of films, music or computer programs	<i>a university library</i> <i>a music library, a software library</i>
Line	Polysemous	A long and thin mark on a surface Row of words which are part of a bigger text	<i>a straight line, a dotted line</i> <i>lines in a play, characters per line</i>
Lunch	Polysemous	Meal eaten in the middle of the day (i.e. the actual food) Time period in which lunch is eaten	<i>a healthy lunch, eating lunch</i> <i>a long lunch, an eventful lunch</i>

Noun	Ambiguity	Definitions	Examples
Message	Polysemous	Short piece of information that you give to a person when you cannot speak to them directly The most important idea in a book, film, or play, or an idea that you want to tell people about	<i>leave a message, send a message</i> <i>the movie's message, getting the message across</i>
Notice	Polysemous	A board or piece of paper containing information or instructions (physical object) Information or a warning given about something that is going to happen in the future	<i>a large notice, a notice in the papers</i> <i>a month's notice, until further notice</i>
Organ	Homonymous	Part of the body of an animal or plant Musical instrument with a keyboard	<i>internal organs, organ transplants</i> <i>playing the organ, pipe organ</i>
Palm	Homonymous	The inside part of your hand A type of tropical tree	<i>sweaty palms, palm reading</i> <i>date palms, palm fronds</i>
Paper	Polysemous	Material used for writing, drawing or printing Document such as a newspaper or piece of writing...	<i>a sheet of paper, wrapping paper, recycled paper</i> <i>the local papers, a scientific paper</i>
Pen	Homonymous	Device used for writing Small area surrounded by a fence	<i>a fountain pen, a ball-point pen</i> <i>a sheep pen</i>
Perch	Homonymous	A type of fish A seat or other place high up	<i>perch recipes, fishing perch</i> <i>birds perch, watching from a perch</i>
Pitcher	Homonymous	Type of jug Baseball player that throws the ball to the batter	<i>a ceramic pitcher, a pitcher of water</i> <i>the pitcher threw a lot of curve balls</i>

Noun	Ambiguity	Definitions	Examples
Plant	Homonymous	Living organism growing in earth Factory or industrial process	<i>a potato plant</i> <i>power plants, water treatment plants</i>
Port	Homonymous	A harbour or the town that has a harbour A strong and sweet wine	<i>a fishing port</i> <i>port wine, a glass of port</i>
Press	Homonymous	Newspapers and magazines, and those parts of television and radio that broadcast news, or reporters and photographers who work for them Piece of equipment that is used to put weight on something in order to crush it, remove liquid from it or to make it flat	<i>press reports, national press</i> <i>a printing press, a garlic press</i>
Punch	Homonymous	A forceful hit with a fist A drink made from fruit juices	<i>a punch on the nose</i> <i>a coconut punch</i>
Shower	Polysemous	Device releasing water that is used to wash oneself Act of washing (using a shower)	<i>a broken shower, a ceiling shower</i> <i>a daily shower, a quick shower</i>
Sign	Polysemous	A notice giving information, directions, a warning, etc. Something showing that something else exists or might happen or exist in the future	<i>a road sign, a neon sign</i> <i>signs of improvement, no sign of life</i>
Straw	Homonymous	Dried stems of crops Thin tube used for drinking	<i>a bale of straw, a straw hat</i> <i>drinking milk through a straw, a paper straw</i>
Swallow	Homonymous	A type of small bird Act of swallowing	<i>swallow migration</i> <i>a swallow of beer</i>

Noun	Ambiguity	Definitions	Examples
Television	Polysemous	A device with a screen (physical object) Broadcasting companies or their programs	<i>a HD television, a new television</i> <i>live television, watching television, on television</i>
Tin	Polysemous	Silver coloured metal / Chemical element Closed metal container in which food is sold	<i>tin mining, tin candlestick</i> <i>tin of beans, soup tins</i>
Trip	Homonymous	Journey in which you go somewhere Occasion when you knock your foot against something and fall or lose your balance	<i>a round trip, a trip to Paris</i> <i>a nasty trip on the stairs</i>
Volume	Polysemous	A book or a book within a series of related books The number or amount of something	<i>volumes of an encyclopedia</i> <i>the volume of traffic, the volume of the music</i>
Wheat	Polysemous	Crop that is used for making flour Food ingredient which has nutritional value	<i>wheat fields, common wheat</i> <i>wheat allergy, wheat products</i>
Yarn	Homonymous	Thread used for making cloth or for knitting A story, usually a long one with a lot of excitement or interest	<i>knitting yarn</i> <i>a boys' adventure yarn</i>

Ambiguous Verbs

Verb	Ambiguity	Definitions	Examples
Admit	Homonymous	To acknowledge that something is true (literal or figurative) Allow someone to enter a place (literal or figurative)	<i>admitting guilt, admitting a mistake</i> <i>admitting new students, admitting to hospital</i>
Adopt	Polysemous	Take a child or a pet into your home (literal only) To choose something/someone as your own, take up a new feature (figurative only)	<i>Adopt a child, adopting a dog</i> <i>adopt an new strategy, adopt a new attitude</i>
Bill	Homonymous	To request payment for a product or service (literal or figurative) to advertise something with a particular description (literal or figurative)	<i>bill the company for your expenses</i> <i>bill the movie as a romantic comedy</i>
Bore	Homonymous	To talk or act in a way that makes someone lose interest (literal or figurative) To make a hole in something (literal or figurative)	<i>boring an audience</i> <i>boring a hole in the wall</i>
Box	Homonymous	To put something in a box (literal or figurative) To practice the sport of boxing (literal or figurative)	<i>boxing up clothes and books</i> <i>He boxed professionally for years</i>
Bury	Polysemous	Put something in the ground (literal only) To hide something (figurative only)	<i>bury a body, bury a treasure</i> <i>bury your face in your hands, bury the truth</i>

Verb	Ambiguity	Definitions	Examples
Capture	Polysemous	Take something into your possession (literal only) Represent or describe something accurately (figurative only)	<i>capturing prisoners, capturing a ball</i> <i>capturing an idea, capturing a picture</i>
Cast	Homonymous	To choose actors for a part in a play or show (literal or figurative) To send something in a particular direction (literal or figurative)	<i>he was cast as the villain</i> <i>casting light, casting shadows</i>
Catch	Polysemous	Take hold of something with your hands (literal only) Engage a person's interest or imagination (figurative only)	<i>catching a ball</i> <i>catching a movie, catching what they said</i>
Charge	Homonymous	To accuse somebody of doing something (usually criminal) (literal or figurative) To put electricity into a device (literal or figurative)	<i>charging someone with murder</i> <i>charging your phone</i>
Clean	Polysemous	Remove dirt or stains (literal only) Remove or eradicate something (figurative only)	<i>cleaning the table</i> <i>clean your thoughts, clean up their behaviour</i>
Climb	Polysemous	Go up towards something (literal only) Increase in scale/value or power (figurative only)	<i>climbing a mountain</i> <i>climb the social ladder, prices climbed</i>
Conduct	Homonymous	Organise or perform an activity (literal or figurative) Allow something through (literal or figurative)	<i>conduct an experiment, conduct a survey</i> <i>conduct electricity, conduct heat</i>
Copy	Polysemous	Produce something so that is the same as an original piece (literal only) Behave the same way as someone/something else (figurative only)	<i>copy documents or data, copy someone's homework</i> <i>copying a role model</i>

Verb	Ambiguity	Definitions	Examples
Dice	Homonymous	To cut in small cubes (literal or figurative) Gamble / Take risks with (literal or figurative)	<i>dicing carrots and potatoes</i> <i>dicing with death</i>
Disarm	Polysemous	Take weapons away from somebody/give up weapons (literal only) To make somebody less angry or critical (figurative only)	<i>disarming rebels, disarming a country</i> <i>disarming critics</i>
File	Homonymous	To store information (literal or figurative) Make an object/surface smooth using a tool (literal or figurative)	<i>filing a document, filing a report</i> <i>filing nails</i>
Follow	Polysemous	To move behind someone or something (literal only) To happen as a result (figurative only)	<i>the cat followed the string</i> <i>following order, the result followed</i>
Grasp	Polysemous	Grab something firmly in your hands (literal only) Understand something (figurative only)	<i>grasping a bottle</i> <i>grasping a concept</i>
Inherit	Polysemous	To receive money or property from a dead relative (literal only) To receive/be left with something a predecessor (figurative only)	<i>inheriting a house, inheriting the family business</i> <i>inherit issues from the previous government</i>
Label	Polysemous	Stick or fasten a piece of paper/fabric containing some information about a product (literal only) Classify something or someone (figurative only)	<i>labelling parcels, labelling jars</i> <i>labelling groups of people</i>

Verb	Ambiguity	Definitions	Examples
Lap	Homonymous	(Usually for animals) Take up (liquid) with quick movements of the tongue (literal or figurative) Complete a full trip around a track or overtake someone in a race by completing one more lap (literal or figurative)	<i>the dog lapped the water, waves lapped the shore</i> <i>lapping a competitor, lapping the track</i>
Launch	Polysemous	Send something out in the air (literal only) Begin something new or launch a new product (figurative only)	<i>launch a projectile, launch a spaceship</i> <i>launch a new brand, launch a programme</i>
Leak	Polysemous	To escape from a hole or crack; allowing liquid or gas to escape (literal only) Intentionally disclosing (private or secret) information (figurative only)	<i>water leaked out of the pipe, the bottle leaked</i> <i>leaking documents</i>
Milk	Polysemous	To get milk from an animal (literal only) Exploit information, money, etc, from something or someone (figurative only)	<i>milking a cow or a goat</i> <i>milking a resource, the media milked the story</i>
Mount	Polysemous	Get on something (literal only) To gradually increase (figurative only)	<i>mounting a horse</i> <i>Tension was mounting</i>
Patronise	Homonymous	Speak or behave in a way that betrays a feeling a superiority (literal or figurative) Be a regular customer in a place (literal or figurative)	<i>patronising kids</i> <i>patronising a bar or a restaurant</i>
Pen	Homonymous	To write something (literal or figurative) Enclose in a restricted space (literal or figurative)	<i>pen a note</i> <i>the farmer penned their sheep</i>

Verb	Ambiguity	Definitions	Examples
Plug	Homonymous	Fill a hole with a piece suitable material (literal or figurative) Advertise/Praising something (literal or figurative)	<i>plugging a charger, plugging a gap</i> <i>plugging a book or a show</i>
Reach	Polysemous	To arrive at a place (literal only) To get to a particular level (figurative only)	<i>reaching a destination, reaching for the phone</i> <i>reaching old age</i>
Reflect	Polysemous	Throw something back without absorbing it (literal only) Think or be representative of something (figurative only)	<i>reflecting light, reflecting waves</i> <i>reflecting about foreign politics, reflecting statistics</i>
Ruin	Polysemous	To cause someone to no longer have money (literal only) Spoil or destroy something (figurative only)	<i>ruining businesses</i> <i>ruining someone's chances, ruining a movie</i>
Rule	Homonymous	Draw a straight line (literal or figurative) Have authority or control over something (literal or figurative)	<i>ruling the paper horizontally</i> <i>ruling a country, ruling the current market</i>
Seize	Polysemous	To take something and keep/hold it (literal only) To take hold of something (figurative only)	<i>seizing one's arm</i> <i>seizing an opportunity</i>
Sketch	Polysemous	Draw something (literal only) To describe something (figurative only)	<i>the artist sketched the model</i> <i>sketching a proof</i>
Stir	Polysemous	To mix a liquid (literal only) Make someone feel a strong emotion (figurative only)	<i>stirring a soup</i> <i>the speech stirred the crowd</i>

Verb	Ambiguity	Definitions	Examples
Tackle	Polysemous	(Sport) Try to take the ball from an opponent; take hold of another player (literal only) To deal with something (figurative only)	<i>The player tackled their opponent</i> <i>tackle a problem</i>
Tap	Homonymous	Hit something gently (literal or figurative) Secretly listen or record what someone is saying using a device (literal or figurative)	<i>tap someone on the shoulder</i> <i>tapping phones</i>
Throw	Polysemous	Send something in the air (literal only) Send something to into a different state (figurative only)	<i>throwing a ball</i> <i>throwing a party, throwing away something, throwing a shadow</i>
Tip	Homonymous	To give an extra amount of money for a service (literal or figurative) To move an object such that one side is higher than the other (literal or figurative)	<i>tip the waiter or taxi driver</i> <i>tipping the table, tip the content of the container</i>
Wipe	Polysemous	Clean a surface by rubbing it with a cloth (literal only) Remove or eliminate completely (figurative only)	<i>wiping the table</i> <i>wiping a hard drive, wiping out an idea</i>

D.2 The corpus dataset

Rank 2 models

Verb	Noun	Verb	Noun	Verb	Noun
admit	band	file	volume	reflect	letter
admit	letter	follow	band	reflect	trip
admit	press	follow	bank	reflect	volume
adopt	band	follow	press	ruin	plant
adopt	bank	follow	trip	saw	cabinet
adopt	boxer	grasp	band	saw	volume
adopt	coach	grasp	palm	scent	plant
bill	band	inherit	bank	seize	bank
bury	boxer	inherit	boxer	seize	press
bury	letter	inherit	plant	sketch	organ
capture	band	label	press	stir	plant
capture	bank	launch	band	tackle	bank
capture	fall	launch	coach	tap	cabinet
capture	organ	launch	letter	tap	pen
capture	palm	launch	port	throw	band
capture	plant	leak	pen	throw	bank
capture	port	mount	band	throw	letter
capture	press	mount	cabinet	throw	pen
cast	band	mount	coach	throw	pitcher
cast	bank	mount	volume	throw	punch
cast	beam	mount	watch	throw	volume
cast	coach	pen	letter	wipe	bank
cast	letter	plug	band	wipe	pen
cast	yarn	reach	band	wipe	plant
catch	letter	reach	bank		
clean	organ	reach	coach		
climb	boxer	reach	pitcher		
climb	palm	reach	plant		
climb	plant	reach	port		
copy	band	reflect	band		
copy	letter	reflect	bank		
file	cabinet	reflect	beam		
file	plant	reflect	coach		

Rank 4 models - Subject-Verb models

Subjects	Verbs	Subjects	Verbs
band, bank	admit, adopt	band, bank	follow, reflect
band, bank	admit, capture	band, bank	follow, seize
band, bank	admit, cast	band, bank	follow, tackle
band, bank	admit, follow	band, bank	follow, throw
band, bank	admit, launch	band, bank	launch, reach
band, bank	admit, reach	band, bank	launch, reflect
band, bank	admit, reflect	band, bank	launch, seize
band, bank	admit, seize	band, bank	launch, tackle
band, bank	admit, tackle	band, bank	launch, throw
band, bank	admit, throw	band, bank	reach, reflect
band, bank	adopt, capture	band, bank	reach, seize
band, bank	adopt, cast	band, bank	reach, tackle
band, bank	adopt, follow	band, bank	reach, throw
band, bank	adopt, launch	band, bank	reflect, seize
band, bank	adopt, reach	band, bank	reflect, tackle
band, bank	adopt, reflect	band, bank	reflect, throw
band, bank	adopt, seize	band, bank	seize, tackle
band, bank	adopt, tackle	band, bank	seize, throw
band, bank	adopt, throw	band, bank	tackle, throw
band, bank	capture, cast	band, beam	cast, reach
band, bank	capture, follow	band, beam	cast, reflect
band, bank	capture, launch	band, beam	reach, reflect
band, bank	capture, reach	band, cabinet	adopt, cast
band, bank	capture, reflect	band, cabinet	adopt, mount
band, bank	capture, seize	band, cabinet	adopt, tap
band, bank	capture, tackle	band, cabinet	cast, mount
band, bank	capture, throw	band, cabinet	cast, tap
band, bank	cast, follow	band, cabinet	mount, tap
band, bank	cast, launch	band, coach	admit, adopt
band, bank	cast, reach	band, coach	admit, cast
band, bank	cast, reflect	band, coach	admit, copy
band, bank	cast, seize	band, coach	admit, launch
band, bank	cast, tackle	band, coach	admit, mount
band, bank	cast, throw	band, coach	admit, reach
band, bank	follow, launch	band, coach	admit, reflect
band, bank	follow, reach	band, coach	admit, tackle

Subjects	Verbs	Subjects	Verbs
band, coach	admit, tap	band, coach	mount, throw
band, coach	admit, throw	band, coach	reach, reflect
band, coach	adopt, cast	band, coach	reach, tackle
band, coach	adopt, copy	band, coach	reach, tap
band, coach	adopt, launch	band, coach	reach, throw
band, coach	adopt, mount	band, coach	reflect, tackle
band, coach	adopt, reach	band, coach	reflect, tap
band, coach	adopt, reflect	band, coach	reflect, throw
band, coach	adopt, tackle	band, coach	tackle, tap
band, coach	adopt, tap	band, coach	tackle, throw
band, coach	adopt, throw	band, coach	tap, throw
band, coach	cast, copy	band, letter	admit, capture
band, coach	cast, launch	band, letter	admit, cast
band, coach	cast, mount	band, letter	admit, copy
band, coach	cast, reach	band, letter	admit, launch
band, coach	cast, reflect	band, letter	admit, pen
band, coach	cast, tackle	band, letter	admit, reach
band, coach	cast, tap	band, letter	admit, reflect
band, coach	cast, throw	band, letter	admit, stir
band, coach	copy, launch	band, letter	admit, throw
band, coach	copy, mount	band, letter	capture, cast
band, coach	copy, reach	band, letter	capture, copy
band, coach	copy, reflect	band, letter	capture, launch
band, coach	copy, tackle	band, letter	capture, pen
band, coach	copy, tap	band, letter	capture, reach
band, coach	copy, throw	band, letter	capture, reflect
band, coach	launch, mount	band, letter	capture, stir
band, coach	launch, reach	band, letter	capture, throw
band, coach	launch, reflect	band, letter	cast, copy
band, coach	launch, tackle	band, letter	cast, launch
band, coach	launch, tap	band, letter	cast, pen
band, coach	launch, throw	band, letter	cast, reach
band, coach	mount, reach	band, letter	cast, reflect
band, coach	mount, reflect	band, letter	cast, stir
band, coach	mount, tackle	band, letter	cast, throw
band, coach	mount, tap	band, letter	copy, launch

Subjects	Verbs	Subjects	Verbs
band, letter	copy, pen	band, plant	capture, stir
band, letter	copy, reach	band, plant	capture, tap
band, letter	copy, reflect	band, plant	cast, plug
band, letter	copy, stir	band, plant	cast, reach
band, letter	copy, throw	band, plant	cast, reflect
band, letter	launch, pen	band, plant	cast, stir
band, letter	launch, reach	band, plant	cast, tap
band, letter	launch, reflect	band, plant	plug, reach
band, letter	launch, stir	band, plant	plug, reflect
band, letter	launch, throw	band, plant	plug, stir
band, letter	pen, reach	band, plant	plug, tap
band, letter	pen, reflect	band, plant	reach, reflect
band, letter	pen, stir	band, plant	reach, stir
band, letter	pen, throw	band, plant	reach, tap
band, letter	reach, reflect	band, plant	reflect, stir
band, letter	reach, stir	band, plant	reflect, tap
band, letter	reach, throw	band, plant	stir, tap
band, letter	reflect, stir	band, port	capture, grasp
band, letter	reflect, throw	band, port	capture, launch
band, letter	stir, throw	band, port	capture, reach
band, palm	capture, grasp	band, port	grasp, launch
band, palm	capture, throw	band, port	grasp, reach
band, palm	grasp, throw	band, port	launch, reach
band, pen	mount, plug	band, press	admit, capture
band, pen	mount, reach	band, press	admit, cast
band, pen	mount, tap	band, press	admit, follow
band, pen	mount, throw	band, press	admit, grasp
band, pen	plug, reach	band, press	admit, reflect
band, pen	plug, tap	band, press	admit, seize
band, pen	plug, throw	band, press	admit, stir
band, pen	reach, tap	band, press	capture, cast
band, pen	reach, throw	band, press	capture, follow
band, pen	tap, throw	band, press	capture, grasp
band, plant	capture, cast	band, press	capture, reflect
band, plant	capture, plug	band, press	capture, seize
band, plant	capture, reach	band, press	capture, stir
band, plant	capture, reflect	band, press	cast, follow

Subjects	Verbs	Subjects	Verbs
band, press	cast, grasp	bank, beam	reach, reflect
band, press	cast, reflect	bank, boxer	adopt, cast
band, press	cast, seize	bank, boxer	adopt, inherit
band, press	cast, stir	bank, boxer	cast, inherit
band, press	follow, grasp	bank, cabinet	adopt, cast
band, press	follow, reflect	bank, cabinet	adopt, file
band, press	follow, seize	bank, cabinet	cast, file
band, press	follow, stir	bank, coach	admit, adopt
band, press	grasp, reflect	bank, coach	admit, cast
band, press	grasp, seize	bank, coach	admit, launch
band, press	grasp, stir	bank, coach	admit, reach
band, press	reflect, seize	bank, coach	admit, reflect
band, press	reflect, stir	bank, coach	admit, tackle
band, press	seize, stir	bank, coach	admit, throw
band, trip	bore, follow	bank, coach	adopt, cast
band, trip	bore, reflect	bank, coach	adopt, launch
band, trip	bore, throw	bank, coach	adopt, reach
band, trip	follow, reflect	bank, coach	adopt, reflect
band, trip	follow, throw	bank, coach	adopt, tackle
band, trip	reflect, throw	bank, coach	adopt, throw
band, volume	capture, mount	bank, coach	cast, launch
band, volume	capture, plug	bank, coach	cast, reach
band, volume	capture, reach	bank, coach	cast, reflect
band, volume	capture, reflect	bank, coach	cast, tackle
band, volume	capture, throw	bank, coach	cast, throw
band, volume	mount, plug	bank, coach	launch, reach
band, volume	mount, reach	bank, coach	launch, reflect
band, volume	mount, reflect	bank, coach	launch, tackle
band, volume	mount, throw	bank, coach	launch, throw
band, volume	plug, reach	bank, coach	reach, reflect
band, volume	plug, reflect	bank, coach	reach, tackle
band, volume	plug, throw	bank, coach	reach, throw
band, volume	reach, reflect	bank, coach	reflect, tackle
band, volume	reach, throw	bank, coach	reflect, throw
band, volume	reflect, throw	bank, coach	tackle, throw
bank, beam	cast, reach	bank, letter	admit, capture
bank, beam	cast, reflect	bank, letter	admit, cast

Subjects	Verbs	Subjects	Verbs
bank, letter	admit, launch	bank, plant	cast, wipe
bank, letter	admit, reach	bank, plant	file, inherit
bank, letter	admit, reflect	bank, plant	file, leak
bank, letter	admit, throw	bank, plant	file, reach
bank, letter	capture, cast	bank, plant	file, reflect
bank, letter	capture, launch	bank, plant	file, wipe
bank, letter	capture, reach	bank, plant	inherit, leak
bank, letter	capture, reflect	bank, plant	inherit, reach
bank, letter	capture, throw	bank, plant	inherit, reflect
bank, letter	cast, launch	bank, plant	inherit, wipe
bank, letter	cast, reach	bank, plant	leak, reach
bank, letter	cast, reflect	bank, plant	leak, reflect
bank, letter	cast, throw	bank, plant	leak, wipe
bank, letter	launch, reach	bank, plant	reach, reflect
bank, letter	launch, reflect	bank, plant	reach, wipe
bank, letter	launch, throw	bank, plant	reflect, wipe
bank, letter	reach, reflect	bank, port	capture, launch
bank, letter	reach, throw	bank, port	capture, reach
bank, letter	reflect, throw	bank, port	launch, reach
bank, pen	leak, reach	bank, press	admit, capture
bank, pen	leak, throw	bank, press	admit, cast
bank, pen	leak, wipe	bank, press	admit, follow
bank, pen	reach, throw	bank, press	admit, lap
bank, pen	reach, wipe	bank, press	admit, leak
bank, pen	throw, wipe	bank, press	admit, reflect
bank, plant	capture, cast	bank, press	admit, seize
bank, plant	capture, file	bank, press	capture, cast
bank, plant	capture, inherit	bank, press	capture, follow
bank, plant	capture, leak	bank, press	capture, lap
bank, plant	capture, reach	bank, press	capture, leak
bank, plant	capture, reflect	bank, press	capture, reflect
bank, plant	capture, wipe	bank, press	capture, seize
bank, plant	cast, file	bank, press	cast, follow
bank, plant	cast, inherit	bank, press	cast, lap
bank, plant	cast, leak	bank, press	cast, leak
bank, plant	cast, reach	bank, press	cast, reflect
bank, plant	cast, reflect	bank, press	cast, seize

Subjects	Verbs	Subjects	Verbs
bank, press	follow, lap	beam, plant	cast, reach
bank, press	follow, leak	beam, plant	cast, reflect
bank, press	follow, reflect	beam, plant	catch, reach
bank, press	follow, seize	beam, plant	catch, reflect
bank, press	lap, leak	beam, plant	reach, reflect
bank, press	lap, reflect	beam, press	cast, catch
bank, press	lap, seize	beam, press	cast, reflect
bank, press	leak, reflect	beam, press	catch, reflect
bank, press	leak, seize	boxer, letter	box, bury
bank, press	reflect, seize	boxer, letter	box, cast
bank, trip	follow, reflect	boxer, letter	bury, cast
bank, trip	follow, throw	boxer, plant	box, cast
bank, trip	follow, wipe	boxer, plant	box, climb
bank, trip	reflect, throw	boxer, plant	box, inherit
bank, trip	reflect, wipe	boxer, plant	cast, climb
bank, trip	throw, wipe	boxer, plant	cast, inherit
bank, volume	capture, file	boxer, plant	climb, inherit
bank, volume	capture, reach	cabinet, coach	adopt, cast
bank, volume	capture, reflect	cabinet, coach	adopt, mount
bank, volume	capture, throw	cabinet, coach	adopt, tap
bank, volume	file, reach	cabinet, coach	cast, mount
bank, volume	file, reflect	cabinet, coach	cast, tap
bank, volume	file, throw	cabinet, coach	mount, tap
bank, volume	reach, reflect	cabinet, plant	cast, file
bank, volume	reach, throw	cabinet, plant	cast, tap
bank, volume	reflect, throw	cabinet, plant	file, tap
bark, plant	catch, climb	cabinet, volume	file, mount
beam, coach	cast, reach	coach, letter	admit, cast
beam, coach	cast, reflect	coach, letter	admit, copy
beam, coach	reach, reflect	coach, letter	admit, launch
beam, letter	cast, catch	coach, letter	admit, reach
beam, letter	cast, reach	coach, letter	admit, reflect
beam, letter	cast, reflect	coach, letter	admit, throw
beam, letter	catch, reach	coach, letter	cast, copy
beam, letter	catch, reflect	coach, letter	cast, launch
beam, letter	reach, reflect	coach, letter	cast, reach
beam, plant	cast, catch	coach, letter	cast, reflect

Subjects	Verbs	Subjects	Verbs
coach, letter	cast, throw	letter, plant	box, cast
coach, letter	copy, launch	letter, plant	box, catch
coach, letter	copy, reach	letter, plant	box, reach
coach, letter	copy, reflect	letter, plant	box, reflect
coach, letter	copy, throw	letter, plant	box, stir
coach, letter	launch, reach	letter, plant	capture, cast
coach, letter	launch, reflect	letter, plant	capture, catch
coach, letter	launch, throw	letter, plant	capture, reach
coach, letter	reach, reflect	letter, plant	capture, reflect
coach, letter	reach, throw	letter, plant	capture, stir
coach, letter	reflect, throw	letter, plant	cast, catch
coach, pen	mount, reach	letter, plant	cast, reach
coach, pen	mount, rule	letter, plant	cast, reflect
coach, pen	mount, tap	letter, plant	cast, stir
coach, pen	mount, throw	letter, plant	catch, reach
coach, pen	reach, rule	letter, plant	catch, reflect
coach, pen	reach, tap	letter, plant	catch, stir
coach, pen	reach, throw	letter, plant	reach, reflect
coach, pen	rule, tap	letter, plant	reach, stir
coach, pen	rule, throw	letter, plant	reflect, stir
coach, pen	tap, throw	letter, port	capture, launch
coach, plant	cast, reach	letter, port	capture, reach
coach, plant	cast, reflect	letter, port	launch, reach
coach, plant	cast, tap	letter, press	admit, capture
coach, plant	reach, reflect	letter, press	admit, cast
coach, plant	reach, tap	letter, press	admit, catch
coach, plant	reflect, tap	letter, press	admit, reflect
coach, press	admit, cast	letter, press	admit, stir
coach, press	admit, reflect	letter, press	capture, cast
coach, press	cast, reflect	letter, press	capture, catch
coach, volume	mount, reach	letter, press	capture, reflect
coach, volume	mount, reflect	letter, press	capture, stir
coach, volume	mount, throw	letter, press	cast, catch
coach, volume	reach, reflect	letter, press	cast, reflect
coach, volume	reach, throw	letter, press	cast, stir
coach, volume	reflect, throw	letter, press	catch, reflect
letter, plant	box, capture	letter, press	catch, stir

Subjects	Verbs	Subjects	Verbs
letter, press	reflect, stir	plant, press	reflect, stir
letter, volume	capture, reach	plant, volume	capture, file
letter, volume	capture, reflect	plant, volume	capture, plug
letter, volume	capture, throw	plant, volume	capture, reach
letter, volume	reach, reflect	plant, volume	capture, reflect
letter, volume	reach, throw	plant, volume	file, plug
letter, volume	reflect, throw	plant, volume	file, reach
pen, plant	leak, plug	plant, volume	file, reflect
pen, plant	leak, reach	plant, volume	plug, reach
pen, plant	leak, tap	plant, volume	plug, reflect
pen, plant	leak, wipe	plant, volume	reach, reflect
pen, plant	plug, reach	plant, volume	reach, reflect
pen, plant	plug, tap		
pen, plant	plug, wipe		
pen, plant	reach, tap		
pen, plant	reach, wipe		
pen, plant	tap, wipe		
pen, volume	mount, plug		
pen, volume	mount, reach		
pen, volume	mount, throw		
pen, volume	plug, reach		
pen, volume	plug, throw		
pen, volume	reach, throw		
plant, press	capture, cast		
plant, press	capture, catch		
plant, press	capture, leak		
plant, press	capture, reflect		
plant, press	capture, stir		
plant, press	cast, catch		
plant, press	cast, leak		
plant, press	cast, reflect		
plant, press	cast, stir		
plant, press	catch, leak		
plant, press	catch, reflect		
plant, press	catch, stir		
plant, press	leak, reflect		
plant, press	leak, stir		

Rank 4 models - Verb-Object models

Verbs	Objects	Verbs	Objects
admit, clean	band, letter	adopt, climb	bank, boxer
admit, clean	band, volume	adopt, climb	bank, plant
admit, clean	letter, volume	adopt, climb	boxer, plant
admit, follow	band, letter	adopt, grasp	band, bank
admit, follow	band, press	adopt, grasp	band, plant
admit, follow	band, volume	adopt, grasp	bank, plant
admit, follow	letter, press	adopt, inherit	bank, boxer
admit, follow	letter, volume	adopt, inherit	bank, plant
admit, follow	press, volume	adopt, inherit	boxer, plant
admit, label	band, letter	adopt, launch	band, coach
admit, label	band, press	adopt, launch	band, plant
admit, label	letter, press	adopt, launch	coach, plant
admit, launch	band, letter	adopt, mount	band, bank
admit, launch	band, volume	adopt, mount	band, coach
admit, launch	letter, volume	adopt, mount	band, plant
admit, mount	band, letter	adopt, mount	bank, coach
admit, mount	band, press	adopt, mount	bank, plant
admit, mount	band, volume	adopt, mount	coach, plant
admit, mount	letter, press	adopt, reach	band, bank
admit, mount	letter, volume	adopt, reach	band, coach
admit, mount	press, volume	adopt, reach	band, plant
admit, reflect	band, letter	adopt, reach	bank, coach
admit, reflect	band, volume	adopt, reach	bank, plant
admit, reflect	letter, volume	adopt, reach	coach, plant
admit, seize	letter, press	adopt, reflect	band, bank
admit, seize	letter, volume	adopt, reflect	band, coach
admit, seize	press, volume	adopt, reflect	bank, coach
admit, throw	band, letter	adopt, throw	band, bank
admit, throw	band, volume	adopt, throw	band, boxer
admit, throw	letter, volume	adopt, throw	band, plant
adopt, capture	band, bank	adopt, throw	bank, boxer
adopt, capture	band, plant	adopt, throw	bank, plant
adopt, capture	bank, plant	adopt, throw	boxer, plant
adopt, cast	band, bank	adopt, wipe	band, bank
adopt, cast	band, coach	adopt, wipe	band, plant
adopt, cast	bank, coach	adopt, wipe	bank, plant

Verbs	Objects	Verbs	Objects
bury, clean	letter, organ	capture, mount	bank, plant
bury, clean	letter, plant	capture, mount	bank, press
bury, clean	organ, plant	capture, mount	plant, press
bury, tap	boxer, letter	capture, reach	band, bank
bury, tap	boxer, organ	capture, reach	band, plant
bury, tap	letter, organ	capture, reach	band, port
bury, throw	boxer, letter	capture, reach	bank, plant
bury, throw	boxer, plant	capture, reach	bank, port
bury, throw	letter, plant	capture, reach	plant, port
capture, clean	band, organ	capture, reflect	band, bank
capture, clean	band, plant	capture, reflect	band, trip
capture, clean	organ, plant	capture, reflect	bank, trip
capture, climb	bank, palm	capture, rule	band, bank
capture, climb	bank, plant	capture, rule	band, port
capture, climb	palm, plant	capture, rule	band, trip
capture, follow	band, bank	capture, rule	bank, port
capture, follow	band, press	capture, rule	bank, trip
capture, follow	band, trip	capture, rule	port, trip
capture, follow	bank, press	capture, seize	bank, plant
capture, follow	bank, trip	capture, seize	bank, port
capture, follow	press, trip	capture, seize	bank, press
capture, grasp	band, bank	capture, seize	plant, port
capture, grasp	band, palm	capture, seize	plant, press
capture, grasp	band, plant	capture, seize	port, press
capture, grasp	bank, palm	capture, stir	bank, palm
capture, grasp	bank, plant	capture, stir	bank, plant
capture, grasp	palm, plant	capture, stir	bank, port
capture, label	band, plant	capture, stir	bank, trip
capture, label	band, press	capture, stir	palm, plant
capture, label	plant, press	capture, stir	palm, port
capture, launch	band, plant	capture, stir	palm, trip
capture, launch	band, port	capture, stir	plant, port
capture, launch	plant, port	capture, stir	plant, trip
capture, mount	band, bank	capture, stir	port, trip
capture, mount	band, plant	capture, tackle	bank, plant
capture, mount	band, press	capture, tackle	bank, trip

Verbs	Objects	Verbs	Objects
capture, tackle	plant, trip	cast, mount	bank, beam
capture, throw	band, bank	cast, mount	bank, coach
capture, throw	band, plant	cast, mount	bank, letter
capture, throw	bank, plant	cast, mount	beam, coach
capture, wipe	band, bank	cast, mount	beam, letter
capture, wipe	band, organ	cast, mount	coach, letter
capture, wipe	band, palm	cast, reach	band, bank
capture, wipe	band, plant	cast, reach	band, coach
capture, wipe	bank, organ	cast, reach	bank, coach
capture, wipe	bank, palm	cast, reflect	band, bank
capture, wipe	bank, plant	cast, reflect	band, beam
capture, wipe	organ, palm	cast, reflect	band, coach
capture, wipe	organ, plant	cast, reflect	band, letter
capture, wipe	palm, plant	cast, reflect	bank, beam
cast, catch	band, coach	cast, reflect	bank, coach
cast, catch	band, letter	cast, reflect	bank, letter
cast, catch	coach, letter	cast, reflect	beam, coach
cast, clean	band, beam	cast, reflect	beam, letter
cast, clean	band, letter	cast, reflect	coach, letter
cast, clean	band, straw	cast, throw	band, bank
cast, clean	beam, letter	cast, throw	band, beam
cast, clean	beam, straw	cast, throw	band, letter
cast, clean	letter, straw	cast, throw	bank, beam
cast, follow	band, bank	cast, throw	bank, letter
cast, follow	band, letter	cast, throw	beam, letter
cast, follow	bank, letter	catch, follow	band, letter
cast, grasp	band, bank	catch, follow	band, trip
cast, grasp	band, straw	catch, follow	letter, trip
cast, grasp	bank, straw	catch, launch	band, coach
cast, launch	band, coach	catch, launch	band, letter
cast, launch	band, letter	catch, launch	coach, letter
cast, launch	coach, letter	catch, mount	band, coach
cast, mount	band, bank	catch, mount	band, letter
cast, mount	band, beam	catch, mount	coach, letter
cast, mount	band, coach	catch, reach	band, coach
cast, mount	band, letter	catch, reach	band, perch

Verbs	Objects	Verbs	Objects
catch, reach	coach, perch	clean, launch	band, pen
catch, reflect	band, coach	clean, launch	band, plant
catch, reflect	band, letter	clean, launch	band, volume
catch, reflect	band, trip	clean, launch	letter, pen
catch, reflect	coach, letter	clean, launch	letter, plant
catch, reflect	coach, trip	clean, launch	letter, volume
catch, reflect	letter, trip	clean, launch	pen, plant
clean, copy	band, letter	clean, launch	pen, volume
clean, copy	band, pen	clean, launch	plant, volume
clean, copy	letter, pen	clean, mount	band, beam
clean, file	cabinet, letter	clean, mount	band, cabinet
clean, file	cabinet, plant	clean, mount	band, letter
clean, file	cabinet, volume	clean, mount	band, plant
clean, file	letter, plant	clean, mount	band, volume
clean, file	letter, volume	clean, mount	beam, cabinet
clean, file	plant, volume	clean, mount	beam, letter
clean, follow	band, letter	clean, mount	beam, plant
clean, follow	band, volume	clean, mount	beam, volume
clean, follow	letter, volume	clean, mount	cabinet, letter
clean, grasp	band, pen	clean, mount	cabinet, plant
clean, grasp	band, plant	clean, mount	cabinet, volume
clean, grasp	band, straw	clean, mount	letter, plant
clean, grasp	pen, plant	clean, mount	letter, volume
clean, grasp	pen, straw	clean, mount	plant, volume
clean, grasp	plant, straw	clean, reflect	band, beam
clean, label	band, cabinet	clean, reflect	band, letter
clean, label	band, letter	clean, reflect	band, volume
clean, label	band, pen	clean, reflect	beam, letter
clean, label	band, plant	clean, reflect	beam, volume
clean, label	cabinet, letter	clean, reflect	letter, volume
clean, label	cabinet, pen	clean, seize	letter, plant
clean, label	cabinet, plant	clean, seize	letter, volume
clean, label	letter, pen	clean, seize	plant, volume
clean, label	letter, plant	clean, sketch	organ, plant
clean, label	pen, plant	clean, sketch	organ, volume
clean, launch	band, letter	clean, sketch	plant, volume

Verbs	Objects	Verbs	Objects
clean, tackle	letter, plant	clean, throw	pen, plant
clean, tackle	letter, volume	clean, throw	pen, volume
clean, tackle	plant, volume	clean, throw	plant, volume
clean, tap	beam, cabinet	clean, wipe	band, organ
clean, tap	beam, letter	clean, wipe	band, pen
clean, tap	beam, organ	clean, wipe	band, plant
clean, tap	beam, pen	clean, wipe	band, volume
clean, tap	beam, volume	clean, wipe	organ, pen
clean, tap	cabinet, letter	clean, wipe	organ, plant
clean, tap	cabinet, organ	clean, wipe	organ, volume
clean, tap	cabinet, pen	clean, wipe	pen, plant
clean, tap	cabinet, volume	clean, wipe	pen, volume
clean, tap	letter, organ	clean, wipe	plant, volume
clean, tap	letter, pen	climb, grasp	bank, palm
clean, tap	letter, volume	climb, grasp	bank, plant
clean, tap	organ, pen	climb, grasp	palm, plant
clean, tap	organ, volume	climb, inherit	bank, boxer
clean, tap	pen, volume	climb, inherit	bank, plant
clean, throw	band, beam	climb, inherit	boxer, plant
clean, throw	band, cabinet	climb, stir	bank, palm
clean, throw	band, letter	climb, stir	bank, plant
clean, throw	band, pen	climb, stir	palm, plant
clean, throw	band, plant	climb, throw	bank, boxer
clean, throw	band, volume	climb, throw	bank, plant
clean, throw	beam, cabinet	climb, throw	boxer, plant
clean, throw	beam, letter	climb, wipe	bank, palm
clean, throw	beam, pen	climb, wipe	bank, plant
clean, throw	beam, plant	climb, wipe	palm, plant
clean, throw	beam, volume	copy, label	band, letter
clean, throw	cabinet, letter	copy, label	band, pen
clean, throw	cabinet, pen	copy, label	letter, pen
clean, throw	cabinet, plant	copy, launch	band, letter
clean, throw	cabinet, volume	copy, launch	band, pen
clean, throw	letter, pen	copy, launch	letter, pen
clean, throw	letter, plant	copy, throw	band, letter
clean, throw	letter, volume	copy, throw	band, pen

Verbs	Objects	Verbs	Objects
copy, throw	letter, pen	follow, mount	band, press
file, label	cabinet, letter	follow, mount	band, volume
file, label	cabinet, plant	follow, mount	bank, letter
file, label	letter, plant	follow, mount	bank, press
file, launch	letter, plant	follow, mount	bank, volume
file, launch	letter, volume	follow, mount	letter, press
file, launch	plant, volume	follow, mount	letter, volume
file, mount	cabinet, letter	follow, mount	press, volume
file, mount	cabinet, plant	follow, reflect	band, bank
file, mount	cabinet, volume	follow, reflect	band, letter
file, mount	letter, plant	follow, reflect	band, trip
file, mount	letter, volume	follow, reflect	band, volume
file, mount	plant, volume	follow, reflect	bank, letter
file, seize	letter, plant	follow, reflect	bank, trip
file, seize	letter, volume	follow, reflect	bank, volume
file, seize	plant, volume	follow, reflect	letter, trip
file, tackle	letter, plant	follow, reflect	letter, volume
file, tackle	letter, volume	follow, reflect	trip, volume
file, tackle	plant, volume	follow, rule	band, bank
file, tap	cabinet, letter	follow, rule	band, trip
file, tap	cabinet, volume	follow, rule	bank, trip
file, tap	letter, volume	follow, seize	bank, letter
file, throw	cabinet, letter	follow, seize	bank, press
file, throw	cabinet, plant	follow, seize	bank, volume
file, throw	cabinet, volume	follow, seize	letter, press
file, throw	letter, plant	follow, seize	letter, volume
file, throw	letter, volume	follow, seize	press, volume
file, throw	plant, volume	follow, tackle	bank, letter
follow, label	band, letter	follow, tackle	bank, trip
follow, label	band, press	follow, tackle	bank, volume
follow, label	letter, press	follow, tackle	letter, trip
follow, launch	band, letter	follow, tackle	letter, volume
follow, launch	band, volume	follow, tackle	trip, volume
follow, launch	letter, volume	follow, throw	band, bank
follow, mount	band, bank	follow, throw	band, letter
follow, mount	band, letter	follow, throw	band, volume

Verbs	Objects	Verbs	Objects
follow, throw	bank, letter	grasp, wipe	pen, plant
follow, throw	bank, volume	inherit, mount	bank, cabinet
follow, throw	letter, volume	inherit, mount	bank, plant
follow, wipe	band, bank	inherit, mount	cabinet, plant
follow, wipe	band, volume	inherit, throw	bank, boxer
follow, wipe	bank, volume	inherit, throw	bank, cabinet
grasp, label	band, pen	inherit, throw	bank, plant
grasp, label	band, plant	inherit, throw	boxer, cabinet
grasp, label	pen, plant	inherit, throw	boxer, plant
grasp, launch	band, pen	inherit, throw	cabinet, plant
grasp, launch	band, plant	label, launch	band, letter
grasp, launch	pen, plant	label, launch	band, pen
grasp, mount	band, bank	label, launch	band, plant
grasp, mount	band, plant	label, launch	letter, pen
grasp, mount	bank, plant	label, launch	letter, plant
grasp, reach	band, bank	label, launch	pen, plant
grasp, reach	band, plant	label, mount	band, cabinet
grasp, reach	bank, plant	label, mount	band, letter
grasp, stir	bank, palm	label, mount	band, plant
grasp, stir	bank, plant	label, mount	band, press
grasp, stir	palm, plant	label, mount	cabinet, letter
grasp, throw	band, bank	label, mount	cabinet, plant
grasp, throw	band, pen	label, mount	cabinet, press
grasp, throw	band, plant	label, mount	letter, plant
grasp, throw	bank, pen	label, mount	letter, press
grasp, throw	bank, plant	label, mount	plant, press
grasp, throw	pen, plant	label, seize	letter, plant
grasp, wipe	band, bank	label, seize	letter, press
grasp, wipe	band, palm	label, seize	plant, press
grasp, wipe	band, pen	label, tap	cabinet, letter
grasp, wipe	band, plant	label, tap	cabinet, pen
grasp, wipe	bank, palm	label, tap	letter, pen
grasp, wipe	bank, pen	label, throw	band, cabinet
grasp, wipe	bank, plant	label, throw	band, letter
grasp, wipe	palm, pen	label, throw	band, pen
grasp, wipe	palm, plant	label, throw	band, plant

Verbs	Objects	Verbs	Objects
label, throw	cabinet, letter	launch, seize	port, volume
label, throw	cabinet, pen	launch, tackle	letter, plant
label, throw	cabinet, plant	launch, tackle	letter, volume
label, throw	letter, pen	launch, tackle	plant, volume
label, throw	letter, plant	launch, tap	letter, pen
label, throw	pen, plant	launch, tap	letter, volume
label, wipe	band, pen	launch, tap	pen, volume
label, wipe	band, plant	launch, throw	band, letter
label, wipe	pen, plant	launch, throw	band, pen
launch, mount	band, coach	launch, throw	band, plant
launch, mount	band, letter	launch, throw	band, volume
launch, mount	band, plant	launch, throw	letter, pen
launch, mount	band, volume	launch, throw	letter, plant
launch, mount	coach, letter	launch, throw	letter, volume
launch, mount	coach, plant	launch, throw	pen, plant
launch, mount	coach, volume	launch, throw	pen, volume
launch, mount	letter, plant	launch, throw	plant, volume
launch, mount	letter, volume	launch, wipe	band, pen
launch, mount	plant, volume	launch, wipe	band, plant
launch, reach	band, coach	launch, wipe	band, volume
launch, reach	band, plant	launch, wipe	pen, plant
launch, reach	band, port	launch, wipe	pen, volume
launch, reach	coach, plant	launch, wipe	plant, volume
launch, reach	coach, port	mount, reach	band, bank
launch, reach	plant, port	mount, reach	band, coach
launch, reflect	band, coach	mount, reach	band, plant
launch, reflect	band, letter	mount, reach	bank, coach
launch, reflect	band, volume	mount, reach	bank, plant
launch, reflect	coach, letter	mount, reach	coach, plant
launch, reflect	coach, volume	mount, reflect	band, bank
launch, reflect	letter, volume	mount, reflect	band, beam
launch, seize	letter, plant	mount, reflect	band, coach
launch, seize	letter, port	mount, reflect	band, letter
launch, seize	letter, volume	mount, reflect	band, volume
launch, seize	plant, port	mount, reflect	bank, beam
launch, seize	plant, volume	mount, reflect	bank, coach

Verbs	Objects	Verbs	Objects
mount, reflect	bank, letter	mount, throw	band, watch
mount, reflect	bank, volume	mount, throw	bank, beam
mount, reflect	beam, coach	mount, throw	bank, cabinet
mount, reflect	beam, letter	mount, throw	bank, letter
mount, reflect	beam, volume	mount, throw	bank, plant
mount, reflect	coach, letter	mount, throw	bank, volume
mount, reflect	coach, volume	mount, throw	bank, watch
mount, reflect	letter, volume	mount, throw	beam, cabinet
mount, seize	bank, letter	mount, throw	beam, letter
mount, seize	bank, plant	mount, throw	beam, plant
mount, seize	bank, press	mount, throw	beam, volume
mount, seize	bank, volume	mount, throw	beam, watch
mount, seize	letter, plant	mount, throw	cabinet, letter
mount, seize	letter, press	mount, throw	cabinet, plant
mount, seize	letter, volume	mount, throw	cabinet, volume
mount, seize	plant, press	mount, throw	cabinet, watch
mount, seize	plant, volume	mount, throw	letter, plant
mount, seize	press, volume	mount, throw	letter, volume
mount, tackle	bank, letter	mount, throw	letter, watch
mount, tackle	bank, plant	mount, throw	plant, volume
mount, tackle	bank, volume	mount, throw	plant, watch
mount, tackle	letter, plant	mount, throw	volume, watch
mount, tackle	letter, volume	mount, wipe	band, bank
mount, tackle	plant, volume	mount, wipe	band, plant
mount, tap	beam, cabinet	mount, wipe	band, volume
mount, tap	beam, letter	mount, wipe	bank, plant
mount, tap	beam, volume	mount, wipe	bank, volume
mount, tap	cabinet, letter	mount, wipe	plant, volume
mount, tap	cabinet, volume	reach, reflect	band, bank
mount, tap	letter, volume	reach, reflect	band, coach
mount, throw	band, bank	reach, reflect	bank, coach
mount, throw	band, beam	reach, rule	band, bank
mount, throw	band, cabinet	reach, rule	band, port
mount, throw	band, letter	reach, rule	bank, port
mount, throw	band, plant	reach, seize	bank, plant
mount, throw	band, volume	reach, seize	bank, port

Verbs	Objects	Verbs	Objects
reach, seize	plant, port	reflect, throw	beam, volume
reach, stir	bank, plant	reflect, throw	letter, volume
reach, stir	bank, port	reflect, wipe	band, bank
reach, stir	plant, port	reflect, wipe	band, volume
reach, throw	band, bank	reflect, wipe	bank, volume
reach, throw	band, pitcher	rule, stir	bank, port
reach, throw	band, plant	rule, stir	bank, trip
reach, throw	bank, pitcher	rule, stir	port, trip
reach, throw	bank, plant	seize, stir	bank, plant
reach, throw	pitcher, plant	seize, stir	bank, port
reach, wipe	band, bank	seize, stir	plant, port
reach, wipe	band, plant	seize, tackle	bank, letter
reach, wipe	bank, plant	seize, tackle	bank, plant
reflect, rule	band, bank	seize, tackle	bank, volume
reflect, rule	band, trip	seize, tackle	letter, plant
reflect, rule	bank, trip	seize, tackle	letter, volume
reflect, seize	bank, letter	seize, tackle	plant, volume
reflect, seize	bank, volume	seize, throw	bank, letter
reflect, seize	letter, volume	seize, throw	bank, plant
reflect, tackle	bank, letter	seize, throw	bank, volume
reflect, tackle	bank, trip	seize, throw	letter, plant
reflect, tackle	bank, volume	seize, throw	letter, volume
reflect, tackle	letter, trip	seize, throw	plant, volume
reflect, tackle	letter, volume	seize, wipe	bank, plant
reflect, tackle	trip, volume	seize, wipe	bank, volume
reflect, tap	beam, letter	seize, wipe	plant, volume
reflect, tap	beam, volume	sketch, wipe	organ, plant
reflect, tap	letter, volume	sketch, wipe	organ, volume
reflect, throw	band, bank	sketch, wipe	plant, volume
reflect, throw	band, beam	stir, tackle	bank, plant
reflect, throw	band, letter	stir, tackle	bank, trip
reflect, throw	band, volume	stir, tackle	plant, trip
reflect, throw	bank, beam	stir, wipe	bank, palm
reflect, throw	bank, letter	stir, wipe	bank, plant
reflect, throw	bank, volume	stir, wipe	palm, plant
reflect, throw	beam, letter	tackle, throw	bank, letter

Verbs	Objects		Verbs	Objects
tackle, throw	bank, plant			
tackle, throw	bank, volume			
tackle, throw	letter, plant			
tackle, throw	letter, volume			
tackle, throw	plant, volume			
tackle, wipe	bank, plant			
tackle, wipe	bank, volume			
tackle, wipe	plant, volume			
tap, throw	beam, boxer			
tap, throw	beam, cabinet			
tap, throw	beam, letter			
tap, throw	beam, pen			
tap, throw	beam, punch			
tap, throw	beam, volume		throw, wipe	band, pen
tap, throw	boxer, cabinet		throw, wipe	band, plant
tap, throw	boxer, letter		throw, wipe	band, volume
tap, throw	boxer, pen		throw, wipe	bank, pen
tap, throw	boxer, punch		throw, wipe	bank, plant
tap, throw	boxer, volume		throw, wipe	bank, volume
tap, throw	cabinet, letter		throw, wipe	pen, plant
tap, throw	cabinet, pen		throw, wipe	pen, volume
tap, throw	cabinet, punch		throw, wipe	plant, volume
tap, throw	cabinet, volume			
tap, throw	letter, pen			
tap, throw	letter, punch			
tap, throw	letter, volume			
tap, throw	pen, punch			
tap, throw	pen, volume			
tap, throw	punch, volume			
tap, wipe	organ, palm			
tap, wipe	organ, pen			
tap, wipe	organ, volume			
tap, wipe	palm, pen			
tap, wipe	palm, volume			
tap, wipe	pen, volume			
throw, wipe	band, bank			

D.3 The human judgment dataset

Rank 2 models

Verb	Noun	Verb	Noun	Verb	Noun
admit	atmosphere	bore	cabinet	label	iron
bore	atmosphere	capture	cabinet	launch	iron
climb	atmosphere	reflect	cabinet	leak	iron
conduct	atmosphere	box	chicken	admit	letter
adopt	band	bury	chicken	bury	letter
bore	band	cast	chicken	conduct	letter
charge	band	charge	chicken	dice	letter
copy	band	leak	chicken	grasp	letter
dice	band	admit	coach	throw	letter
label	band	bill	coach	admit	library
launch	band	catch	coach	bury	library
milk	band	charge	coach	pen	library
pen	band	clean	coach	rule	library
rule	band	file	coach	tap	library
tackle	band	throw	coach	admit	line
wipe	band	admit	coat	box	line
bore	bank	bury	coat	bury	line
label	bank	throw	coat	cast	line
reflect	bark	charge	cotton	clean	line
bury	beam	climb	cotton	conduct	line
inherit	beam	inherit	cotton	reflect	line
throw	beam	lap	cotton	wipe	line
file	book	throw	cotton	bore	lunch
inherit	book	admit	fall	dice	lunch
launch	book	conduct	fall	label	lunch
bore	boxer	admit	film	throw	lunch
box	boxer	bill	film	charge	message
capture	boxer	bore	film	clean	message
catch	boxer	cast	film	grasp	message
dice	boxer	launch	film	file	notice
disarm	boxer	rule	film	admit	organ
launch	boxer	reflect	glasses	bill	organ
pen	boxer	admit	iron	box	organ
reflect	boxer	charge	iron	cast	organ
throw	boxer	clean	iron	grasp	organ
tip	boxer	copy	iron	lap	organ

Verb	Noun	Verb	Noun	Verb	Noun
pen	organ	bore	plant		
reflect	organ	conduct	plant		
rule	organ	file	plant		
capture	palm	launch	plant		
catch	palm	pen	plant		
lap	palm	tap	plant		
launch	palm	bill	port		
pen	palm	clean	port	Verb	Noun
bill	paper	lap	port	wipe	television
bore	paper	pen	port	bury	tin
capture	paper	admit	press	clean	tin
charge	paper	box	press	dice	tin
conduct	paper	bury	press	launch	tin
disarm	paper	capture	press	plug	tin
file	paper	conduct	press	throw	tin
grasp	paper	file	press	capture	trip
launch	paper	label	press	conduct	trip
rule	paper	wipe	press	box	volume
tap	paper	clean	punch	cast	volume
bore	pen	leak	punch	conduct	volume
bury	pen	bury	shower	file	volume
clean	pen	label	shower	grasp	volume
climb	pen	capture	sign	patronise	volume
launch	pen	inherit	sign	reflect	volume
rule	pen	launch	sign	tap	volume
tap	pen	dice	straw	catch	wheat
capture	perch	inherit	straw	label	wheat
dice	perch	wipe	straw	throw	wheat
file	perch	admit	swallow		
label	perch	bury	swallow		
pen	perch	catch	swallow		
rule	perch	admit	television		
bury	pitcher	bill	television		
catch	pitcher	box	television		
admit	plant	climb	television		
adopt	plant	label	television		

Rank 4 models

Subjects	Verbs	Subjects	Verbs
atmosphere, fall	admit, conduct	band, paper	charge, launch
atmosphere, film	admit, bore	band, paper	charge, rule
atmosphere, letter	admit, conduct	band, paper	launch, rule
atmosphere, line	admit, conduct	band, pen	bore, launch
atmosphere, paper	bore, conduct	band, pen	bore, rule
atmosphere, pen	bore, climb	band, pen	launch, rule
atmosphere, plant	admit, bore	band, perch	dice, label
atmosphere, plant	admit, conduct	band, perch	dice, pen
atmosphere, plant	bore, conduct	band, perch	dice, rule
atmosphere, press	admit, conduct	band, perch	label, pen
atmosphere, television	admit, climb	band, perch	label, rule
band, bank	bore, label	band, perch	pen, rule
band, boxer	bore, dice	band, plant	adopt, bore
band, boxer	bore, launch	band, plant	adopt, launch
band, boxer	bore, pen	band, plant	adopt, pen
band, boxer	dice, launch	band, plant	bore, launch
band, boxer	dice, pen	band, plant	bore, pen
band, boxer	launch, pen	band, plant	launch, pen
band, film	bore, launch	band, press	label, wipe
band, film	bore, rule	band, straw	dice, wipe
band, film	launch, rule	band, television	label, wipe
band, iron	charge, copy	band, tin	dice, launch
band, iron	charge, label	bank, lunch	bore, label
band, iron	charge, launch	beam, coat	bury, throw
band, iron	copy, label	beam, cotton	inherit, throw
band, iron	copy, launch	beam, letter	bury, throw
band, iron	label, launch	beam, tin	bury, throw
band, library	pen, rule	book, paper	file, launch
band, lunch	bore, dice	book, plant	file, launch
band, lunch	bore, label	book, sign	inherit, launch
band, lunch	dice, label	boxer, cabinet	bore, capture
band, organ	pen, rule	boxer, cabinet	bore, reflect
band, palm	launch, pen	boxer, cabinet	capture, reflect
band, paper	bore, charge	boxer, coach	catch, throw
band, paper	bore, launch	boxer, film	bore, launch
band, paper	bore, rule	boxer, letter	dice, throw

Subjects	Verbs	Subjects	Verbs
boxer, line	box, reflect	chicken, line	box, cast
boxer, lunch	bore, dice	chicken, line	bury, cast
boxer, lunch	bore, throw	chicken, organ	box, cast
boxer, lunch	dice, throw	chicken, press	box, bury
boxer, organ	box, pen	chicken, volume	box, cast
boxer, organ	box, reflect	coach, coat	admit, throw
boxer, organ	pen, reflect	coach, cotton	charge, throw
boxer, palm	capture, catch	coach, film	admit, bill
boxer, palm	capture, launch	coach, iron	admit, charge
boxer, palm	capture, pen	coach, iron	admit, clean
boxer, palm	catch, launch	coach, iron	charge, clean
boxer, palm	catch, pen	coach, letter	admit, throw
boxer, palm	launch, pen	coach, line	admit, clean
boxer, paper	bore, capture	coach, message	charge, clean
boxer, paper	bore, disarm	coach, organ	admit, bill
boxer, paper	bore, launch	coach, paper	bill, charge
boxer, paper	capture, disarm	coach, paper	bill, file
boxer, paper	capture, launch	coach, paper	charge, file
boxer, paper	disarm, launch	coach, plant	admit, file
boxer, pen	bore, launch	coach, port	bill, clean
boxer, perch	capture, dice	coach, press	admit, file
boxer, perch	capture, pen	coach, swallow	admit, catch
boxer, perch	dice, pen	coach, television	admit, bill
boxer, plant	bore, launch	coach, tin	clean, throw
boxer, plant	bore, pen	coach, wheat	catch, throw
boxer, plant	launch, pen	coat, letter	admit, bury
boxer, press	box, capture	coat, letter	admit, throw
boxer, sign	capture, launch	coat, letter	bury, throw
boxer, tin	dice, launch	coat, library	admit, bury
boxer, tin	dice, throw	coat, line	admit, bury
boxer, tin	launch, throw	coat, press	admit, bury
boxer, volume	box, reflect	coat, swallow	admit, bury
boxer, wheat	catch, throw	coat, tin	bury, throw
cabinet, paper	bore, capture	fall, letter	admit, conduct
chicken, iron	charge, leak	fall, line	admit, conduct
chicken, line	box, bury	fall, plant	admit, conduct

Subjects	Verbs	Subjects	Verbs
fall, press	admit, conduct	letter, lunch	dice, throw
film, iron	admit, launch	letter, organ	admit, grasp
film, library	admit, rule	letter, paper	conduct, grasp
film, line	admit, cast	letter, plant	admit, conduct
film, organ	admit, bill	letter, press	admit, bury
film, organ	admit, cast	letter, press	admit, conduct
film, organ	admit, rule	letter, press	bury, conduct
film, organ	bill, cast	letter, swallow	admit, bury
film, organ	bill, rule	letter, tin	bury, dice
film, organ	cast, rule	letter, tin	bury, throw
film, paper	bill, bore	letter, tin	dice, throw
film, paper	bill, launch	letter, volume	conduct, grasp
film, paper	bill, rule	library, line	admit, bury
film, paper	bore, launch	library, organ	admit, pen
film, paper	bore, rule	library, organ	admit, rule
film, paper	launch, rule	library, organ	pen, rule
film, pen	bore, launch	library, paper	rule, tap
film, pen	bore, rule	library, pen	bury, rule
film, pen	launch, rule	library, pen	bury, tap
film, plant	admit, bore	library, pen	rule, tap
film, plant	admit, launch	library, perch	pen, rule
film, plant	bore, launch	library, plant	admit, pen
film, television	admit, bill	library, plant	admit, tap
iron, line	admit, clean	library, plant	pen, tap
iron, message	charge, clean	library, press	admit, bury
iron, paper	charge, launch	library, swallow	admit, bury
iron, pen	clean, launch	line, organ	admit, box
iron, plant	admit, launch	line, organ	admit, cast
iron, press	admit, label	line, organ	admit, reflect
iron, punch	clean, leak	line, organ	box, cast
iron, television	admit, label	line, organ	box, reflect
iron, tin	clean, launch	line, organ	cast, reflect
letter, library	admit, bury	line, pen	bury, clean
letter, line	admit, bury	line, plant	admit, conduct
letter, line	admit, conduct	line, press	admit, box
letter, line	bury, conduct	line, press	admit, bury

Subjects	Verbs	Subjects	Verbs
line, press	admit, conduct	organ, volume	box, cast
line, press	admit, wipe	organ, volume	box, grasp
line, press	box, bury	organ, volume	box, reflect
line, press	box, conduct	organ, volume	cast, grasp
line, press	box, wipe	organ, volume	cast, reflect
line, press	bury, conduct	organ, volume	grasp, reflect
line, press	bury, wipe	palm, paper	capture, launch
line, press	conduct, wipe	palm, perch	capture, pen
line, swallow	admit, bury	palm, plant	launch, pen
line, television	admit, box	palm, port	lap, pen
line, television	admit, wipe	palm, sign	capture, launch
line, television	box, wipe	paper, pen	bore, launch
line, tin	bury, clean	paper, pen	bore, rule
line, volume	box, cast	paper, pen	bore, tap
line, volume	box, conduct	paper, pen	launch, rule
line, volume	box, reflect	paper, pen	launch, tap
line, volume	cast, conduct	paper, pen	rule, tap
line, volume	cast, reflect	paper, perch	capture, file
line, volume	conduct, reflect	paper, perch	capture, rule
lunch, perch	dice, label	paper, perch	file, rule
lunch, tin	dice, throw	paper, plant	bore, conduct
lunch, wheat	label, throw	paper, plant	bore, file
message, paper	charge, grasp	paper, plant	bore, launch
organ, palm	lap, pen	paper, plant	bore, tap
organ, paper	bill, grasp	paper, plant	conduct, file
organ, paper	bill, rule	paper, plant	conduct, launch
organ, paper	grasp, rule	paper, plant	conduct, tap
organ, perch	pen, rule	paper, plant	file, launch
organ, plant	admit, pen	paper, plant	file, tap
organ, port	bill, lap	paper, plant	launch, tap
organ, port	bill, pen	paper, press	capture, conduct
organ, port	lap, pen	paper, press	capture, file
organ, press	admit, box	paper, press	conduct, file
organ, television	admit, bill	paper, sign	capture, launch
organ, television	admit, box	paper, trip	capture, conduct
organ, television	bill, box	paper, volume	conduct, file

Subjects	Verbs
paper, volume	conduct, grasp
paper, volume	conduct, tap
paper, volume	file, grasp
paper, volume	file, tap
paper, volume	grasp, tap
pen, plant	bore, launch
pen, plant	bore, tap
pen, plant	launch, tap
pen, tin	bury, clean
pen, tin	bury, launch
pen, tin	clean, launch
perch, plant	file, pen
perch, press	capture, file
perch, press	capture, label
perch, press	file, label
pitcher, swallow	bury, catch
plant, press	admit, conduct
plant, press	admit, file
plant, press	conduct, file
plant, volume	conduct, file
plant, volume	conduct, tap
plant, volume	file, tap
press, shower	bury, label
press, swallow	admit, bury
press, television	admit, box
press, television	admit, label
press, television	admit, wipe
press, television	box, label
press, television	box, wipe
press, television	label, wipe
press, trip	capture, conduct
press, volume	box, conduct
press, volume	box, file
press, volume	conduct, file

We here only listed the SV models. The VO models can be obtained by taking the same verbs, and switching the role of the subjects to objects.

D.4 Prediction dataset

As for the human judgment dataset, we only quote the SV empirical models; the analogue VO models can be obtained by switching the subject to object roles.

Training dataset

Subjects	Verbs	Subjects	Verbs
atmosphere, plant	admit, conduct	boxer, perch	capture, pen
atmosphere, plant	bore, conduct	boxer, perch	dice, pen
band, boxer	launch, pen	boxer, plant	bore, launch
band, film	bore, rule	boxer, plant	launch, pen
band, iron	copy, label	boxer, tin	dice, launch
band, lunch	bore, label	boxer, tin	launch, throw
band, paper	bore, rule	cabinet, paper	bore, capture
band, paper	charge, launch	chicken, line	box, bury
band, paper	launch, rule	coach, coat	admit, throw
band, pen	bore, launch	coach, iron	admit, clean
band, perch	dice, label	coach, message	charge, clean
band, plant	adopt, launch	coach, plant	admit, file
band, plant	bore, pen	coach, television	admit, bill
band, press	label, wipe	coat, letter	admit, bury
beam, tin	bury, throw	film, organ	admit, cast
book, paper	file, launch	film, paper	bore, launch
boxer, lunch	bore, dice	film, pen	launch, rule
boxer, organ	pen, reflect	film, plant	admit, bore
boxer, palm	capture, launch	film, plant	admit, launch
boxer, palm	catch, pen	iron, tin	clean, launch
boxer, paper	bore, disarm	letter, line	admit, bury

Subjects	Verbs	Subjects	Verbs
letter, plant	admit, conduct	palm, port	lap, pen
letter, press	bury, conduct	paper, pen	bore, tap
letter, tin	bury, dice	paper, pen	launch, tap
library, organ	admit, rule	paper, pen	rule, tap
library, organ	pen, rule	paper, perch	capture, file
library, pen	bury, tap	paper, plant	conduct, launch
line, organ	admit, box	paper, press	capture, conduct
line, organ	box, cast	paper, volume	conduct, tap
line, press	box, wipe	paper, volume	file, grasp
line, press	bury, conduct	paper, volume	file, tap
line, television	box, wipe	pen, plant	bore, launch
line, volume	box, cast	pen, tin	bury, clean
line, volume	box, conduct	perch, press	capture, label
line, volume	cast, reflect	plant, press	admit, conduct
lunch, perch	dice, label	plant, volume	conduct, file
lunch, tin	dice, throw	plant, volume	file, tap
organ, paper	bill, rule	press, television	admit, wipe
organ, port	bill, lap	press, volume	box, conduct
organ, television	bill, box		
organ, volume	box, cast		
organ, volume	box, reflect		

Testing dataset

Subjects	Verbs	Subjects	Verbs
organ, television	admit, bill	plant, press	conduct, file
letter, line	admit, conduct	line, television	admit, box
pen, tin	clean, launch	line, volume	box, reflect
line, press	admit, wipe	boxer, palm	launch, pen
line, press	admit, bury	plant, press	admit, file
band, lunch	dice, label	film, paper	bore, rule
boxer, tin	dice, throw	press, television	admit, box
line, press	admit, conduct	paper, plant	file, tap
line, organ	box, reflect	paper, plant	file, launch
line, press	admit, box	line, organ	cast, reflect
band, lunch	bore, dice	film, organ	admit, bill
line, television	admit, wipe	organ, volume	cast, reflect
organ, television	admit, box	perch, press	capture, file
organ, port	lap, pen	pen, plant	launch, tap
paper, volume	conduct, file	paper, plant	conduct, tap
paper, press	capture, file	band, plant	bore, launch
boxer, paper	bore, launch	paper, press	conduct, file
boxer, organ	box, reflect	letter, line	bury, conduct
boxer, paper	bore, capture	band, perch	dice, pen
boxer, lunch	dice, throw	band, plant	launch, pen
boxer, palm	capture, pen	line, press	box, conduct

Subjects	Verbs	Subjects	Verbs
paper, pen	bore, launch	band, boxer	dice, launch
line, press	box, bury	paper, plant	bore, conduct
paper, pen	launch, rule	boxer, plant	bore, pen
paper, plant	bore, launch	film, paper	launch, rule
pen, plant	bore, tap	library, pen	rule, tap
paper, plant	bore, tap	band, boxer	dice, pen
film, organ	bill, rule	atmosphere, plant	admit, bore
coach, iron	charge, clean	coat, letter	admit, throw
boxer, paper	capture, launch	letter, tin	dice, throw
press, volume	conduct, file	film, plant	bore, launch
band, boxer	bore, launch	letter, tin	bury, throw
band, iron	charge, launch	letter, press	admit, bury
film, paper	bill, rule	letter, press	admit, conduct
paper, pen	bore, rule	band, pen	bore, rule
plant, volume	conduct, tap	band, pen	launch, rule
paper, plant	launch, tap	coat, letter	bury, throw
press, television	label, wipe	band, perch	pen, rule
band, boxer	bore, pen	film, pen	bore, launch
band, boxer	bore, dice	film, pen	bore, rule
film, organ	admit, rule	band, film	bore, launch
paper, plant	conduct, file	band, film	launch, rule