Integrating Across Conceptual Spaces

Kaarina Aho

University College London

Department of Experimental Psychology

Submitted to University College London (UCL) in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy.

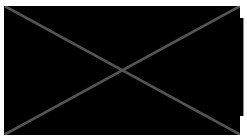
Primary supervisor: Prof. Bradley C. Love

Secondary supervisor: Prof. Gabriella Vigliocco

Thesis submission date: 22/08/2023

Declaration

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



Kaarina Aho

London, United Kingdom

22/08/2023

Abstract

It has been shown that structure is shared across multiple modalities in the real world: if we speak about two items in similar ways, then they are also likely to appear in similar visual contexts. Such similarity relationships are recapitulated across modalities for entire systems of concepts. This provides a signal that can be used to identify the correct mapping between modalities without relying on event-based learning, by a process of *systems alignment*. Because it depends on relationships within a modality, systems alignment can operate *asynchronously*, meaning that learning may not require direct labelling events (e.g., seeing a truck and hearing someone say the word 'truck'). Instead, learning can occur based on linguistic and visual information which is received at different points in time (e.g., having overheard a conversation about trucks, and seeing one on the road the next day).

This thesis explores the value of alignment in learning to integrate between conceptual systems. It takes a joint experimental and computational approach, which simultaneously facilitates insights on alignment processes in controlled environments and at scale.

The role of alignment in learning is explored from three perspectives, yielding three distinct contributions. In Chapter 2, signatures of alignment are identified in a real-world setting: children's early concept learning. Moving to a controlled experimental setting, Chapter 3 demonstrates that humans benefit from alignment signals in cross-system learning, and finds that models which attempt the asynchronous alignment of systems best capture human behaviour. Chapter 4 implements these insights in machine-learning systems, using alignment to tackle cross-modal learning problems at scale.

Alignment processes prove valuable to human learning across conceptual systems, providing a fresh perspective on learning that complements prevailing event-based accounts. This research opens doors for machine learning systems to harness alignment mechanisms for cross-modal learning, thus reducing their reliance on extensive supervision by drawing inspiration from both human learning and the structure of the environment.

Impact Statement

Understanding how humans are able to acquire concept knowledge from their constant and noisy multimodal inputs is of interest both theoretically and practically.

In Chapter 2 of this thesis, I explore a specific real-world alignment opportunity - early concept learning - in detail. It is shown that early acquired concepts are well-positioned to facilitate learning by aligning systems across modalities. This provides a novel insight into how infants, as naive learners, are able to learn correspondences between language and the visual world so successfully from relatively little supervised input. This work sets the stage for future work exploring alignment-based learning in children through behavioural studies.

Within this chapter, I also demonstrate an application of this finding to machine learning systems, by using the structural features associated with alignment to build generative agents which optimise knowledge states for alignment-based learning. The successful application of child-inspired systems contributes to a growing literature on building human-like AI systems, and on optimal curricula for machine systems in the effort to develop more human-like representations of the world.

Chapter 3 of this thesis presents the novel finding that alignable systems contribute to successful learning in humans, even when full supervision is available. In turn, I present computational modelling which suggests that an asynchronous and unsupervised alignment mechanism could underpin this learning behaviour. Given recent findings that systems derived from naturalistic unimodal inputs are alignable across modalities, these insights into human learn-

ing deepen our understanding of how humans learn so successfully in the real world, where supervised learning events are rare and ambiguous.

In this thesis' final contribution in Chapter 4, steps are made towards identifying an algorithm which is able to learn an unsupervised mapping between modalities. First, modifications of the alignment scoring metric are assessed in efforts to find a metric which offers the best chance of algorithmic success. Next, algorithms drawing from a range of related research domains are tested on the problem of unsupervised alignment at various scales. Finally, the impact of alignment-based priors on image classification performance in low data environments is explored.

This thesis contributes to a new perspective on human learning, by demonstrating that human learning is supported by the asynchronous process of aligning across systems. Furthermore, it provides evidence that alignable signals are present when humans begin forming their multimodal understanding of the world. Applying these insights from human learning, it demonstrates the potential for human-inspired machine learning systems to use alignment to capitalise on the rich structural information that is shared across modalities.

For my family, whose boundless support made this possible.

Acknowledgements

One could reasonably expect a PhD conducted over a global pandemic to be an isolating experience, but I am fortunate that my research communities and support networks ensured I never felt alone. I am profoundly grateful to the many people who have shaped my research and personal development throughout this endeavour.

Firstly, I extend my heartfelt gratitude to my supervisor, Professor Brad Love. Brad's kind and astute guidance has had immeasurable impact both on the research presented in this thesis and on my broader understanding of the scientific process and the pursuit of quality research. I thank him for his patience, his ability to infuse enthusiasm into the most challenging phases of research, and for always being available to discuss ideas or provide insightful feedback. This thesis would certainly not have been possible without Brad and his deep and nuanced understanding of the field. I feel privileged to have had his leadership and support.

I am also immensely grateful to Dr. Brett Roads, whose exceptional mentorship and collaborative spirit have been invaluable in this process. Brett's advice, empathy and expertise have gotten me out of ruts on countless occasions. I consider myself extremely fortunate to have worked with Brett, and cannot thank him enough for the extensive time and care that he has generously invested in sharing his knowledge over these years.

I have crossed paths with many other wonderful members of the Love Lab - Xiaoliang Luo, Adam Hornsby, Franziska Bröker, Nikolay Dagaev, Daniel Barry, Sebastian Bobadilla-Suarez, Nick Sexton, and Rob Mok, among others - to whom I am grateful, both for their valuable feedback and for leading by

example in setting high standards of research. Thanks also to the talented MSc students I worked with - Stefania Preda and Guglielmo Reggio - whose enthusiasm and fresh perspectives were inspiring.

I am lucky to have been a part of the Ecological Brain DTP community. Many thanks to the Leverhulme Trust for facilitating the programme, and giving me the opportunity to undertake this PhD. Thanks to my EcoBrain peers, especially my cohort - Nicole Engeler, Viktor Kewenig, Pete Kirk, Hope Oloye, and Leon Reicherts - for their camaraderie and wisdom. Thanks to Professor Mirco Musolesi and Professor Daniel Richardson, for kindly facilitating rotation projects in their labs, where I learned so much. Thanks also to Warda Sharif for being such a reliable and friendly presence.

Thanks in particular are owed to Professor Gabriella Vigliocco. As the EcoBrain director, she has built an inspiring community of multidisciplinary researchers. As my second supervisor on this thesis, she has offered perspectives which have profoundly improved the work, and encouragement which has developed my confidence as a researcher.

Thanks to Professor Chen Yu for sharing his insights and knowledge as a member of my thesis committee. Every meeting with Chen generated exciting ideas, and I am grateful for his enthusiasm for this work.

Beyond the world of academia, I am grateful for my friends and loved ones, who have served as pillars of support and beacons of joy. Above all, I thank my family, for whom my gratitude could not be summarised in a mere 227 pages. Thank you to my sister Erin, for listening to my programming woes and for inspiring me with your intelligence and tenacity. And thank you to my parents, Beth and Aarne, for instilling in me the desire to push my boundaries, and giving me the outlook necessary to embrace this challenge without fear of failure. For your endless generosity, for providing solace on tough days, for believing that I could do this. Thank you, thank you, thank you.

Published articles

Chapters 2 and 3 of this thesis are fully or partially based on the following articles:

- Chapter 2: Aho, K., Roads, B., & Love, B. C. (2023). Signatures of cross-modal alignment in children's early concepts. *PsyArxiv*. `https://psyarxiv.com/v6fnq/`
- Chapter 3: Aho, K., Roads, B. D., & Love, B. C. (2022). System alignment supports cross-domain learning and zero-shot generalisation. *Cognition*, 227, 105200.

I, Kaarina Aho, declare that I was fully involved through the entire process of investigation for these articles, including (but not limited to): experimental design, analysis, and writing of the manuscripts. Declaration forms for the use of these published works are available on the following pages.

UCL Research Paper Declaration Form (#1)

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?** –

B **Please include a link to or doi for the work:** –

C **Where was the work published?** –

D **Who published the work?** –

E **When was the work published?** –

F **List the manuscript's authors in the order they appear on the publication:** –

G **Was the work peer reviewed?** –

H **Have you retained the copyright?** –

I **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes, please give a link or doi** –
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?** Signatures of cross-modal alignment in children's early concepts

B **Has the manuscript been uploaded to a preprint server e.g. 'medRxiv'?** Yes

If 'Yes', please please give a link or doi: https://psyarxiv.com/v6fnq/

C **Where is the work intended to be published?** PNAS

D **List the manuscript's authors in the intended authorship order:** Kaarina Aho, Brett D. Roads, Bradley C. Love

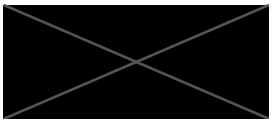
E **Stage of publication:** Second revision under review

3. For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4)**:**
KA: Methodology, Investigation, Formal analysis, Software, Visualization, Preparation of original draft, Approval of final manuscript for submission. BDR: Supervision, Provided revisions for draft, Approval of final manuscript for submission. BCL: Conceptualization, Funding acquisition, Supervision, Provided revisions for draft, Approval of final manuscript for submission.

4. In which chapter(s) of your thesis can this material be found? Chapter 2

e-Signatures confirming that the information above is accurate :

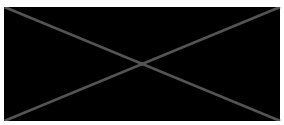
Candidate:



Kaarina Aho

22/08/2023

Supervisor/Senior Author signature:



Bradley C. Love

22/08/2023

UCL Research Paper Declaration Form (#2)

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?**
System alignment supports cross-domain learning and zero-shot generalisation

B **Please include a link to or doi for the work:**
`https://doi.org/10.1016/j.cognition.2022.105200`

C **Where was the work published?** Cognition

D **Who published the work?** Elsevier

E **When was the work published?** October 2022

F **List the manuscript's authors in the order they appear on the publication:** Kaarina Aho, Brett D. Roads, Bradley C. Love

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)?** Yes
If 'Yes', please give a link or doi `https://psyarxiv.com/pm3ay/`
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)**:**

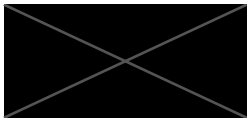
KA: Methodology, Investigation, Formal analysis, Software, Visualization, Preparation of original draft, Approval of final manuscript for submission. BDR: Supervision, Provided revisions for draft, Approval of final manuscript for submission. BCL: Conceptualization, Funding acquisition, Supervision, Provided revisions for draft, Approval of final manuscript for submission.

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

Chapter 3

e-Signatures confirming that the information above is accurate :

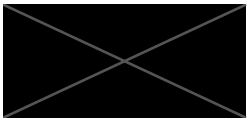
Candidate:



Kaarina Aho

22/08/2023

Supervisor/Senior Author signature:



Bradley C. Love

22/08/2023

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The human experience is inherently multimodal. We are constantly subjected to streams of visual, auditory, haptic, and olfactory information, generated by our environments and the entities contained therein. Further to sensory modalities, language has developed as a mode of communication between humans, allowing us to refer to the entities in the world. These streams of input are noisy, and very challenging to derive meaningful information from as a naive learner. This challenge is evident in the struggles to get machine learning (ML) systems to learn from unstructured, naturalistic data.

And yet, over time and across our experiences, humans integrate this information into a holistic understanding of their surroundings. Experiences of entities are gradually abstracted to become untethered from any individual event, time or place (Lambon Ralph, 2014; Martin, 2016). This forms our semantic knowledge, which in turn allows us to impose structure on our sensory inputs, to generalise to new examples and to make meaningful predictions about the entities in the world around us (Ralph et al., 2017a). On a walk through town, most humans would not be catastrophically overwhelmed by the sounds, sights and smells around them. Effortlessly, they would recognise the whooshing sound accompanied by the sound of rubber rolling over tarmac as being associated with a 'car', and thus would not be surprised when the associated large metal object on four wheels rounded the corner.

Sensory inputs across modalities are fundamentally constrained by a shared

underlying reality. For something to be a 'car', it likely possesses certain properties: it probably has four wheels, with tyres made of some form of rubber; it falls within a certain size range (constrained by the requirement to fit people inside of it, but also to fit on a road and inside of a garage); its body is likely made of metal, and contains a steering wheel and some seats. These properties in turn determine the sensory impact of the car across modalities - what it sounds like, looks like, even smells and tastes like. Their properties are also related to the scenes in which cars will be observed, and the contexts in which they will be spoken about. And all of this contributes to what it means for something to be a car.

A consequence of this is that if two entities are experienced similarly in one modality, they are likely to also be experienced similarly in others. We would likely see a 'cow' and a 'horse' in more similar visual contexts than we would a 'cow' and a 'car', and we would also talk about 'cows' and 'horses' in more similar linguistic contexts than we would 'cows' and 'cars'. The same would be true of the sounds, smells and even visual properties associated with these entities (Johns and Jones, 2012; Roads and Love, 2020). This results in shared structure across modalities in the real world. The existence of shared similarity relationships across modalities, where representations share no physical similarity, is an example of second-order isomorphism (Shepard and Chipman, 1970).

In theory, this shared structure could provide a valuable signal for learning mappings between modalities. If similarity structures are sufficiently replicated across linguistic and visual systems, for example, it could be possible to label all visual items appropriately by aligning the sets of similarity relationships within the modalities. This means that learning could occur in a completely unsupervised fashion, without ever telling the learner which item in one modality corresponds to which item in the other. Consequently, a mapping between modalities could be learned without ever experiencing the two modalities concurrently - that is, via completely asynchronous cross-modal

learning. We call this process of using the idiosyncratic similarity relations that are mirrored across multiple systems to perform a cross-system mapping *systems alignment*. A visualisation of systems alignment is provided in Figure 1.1.

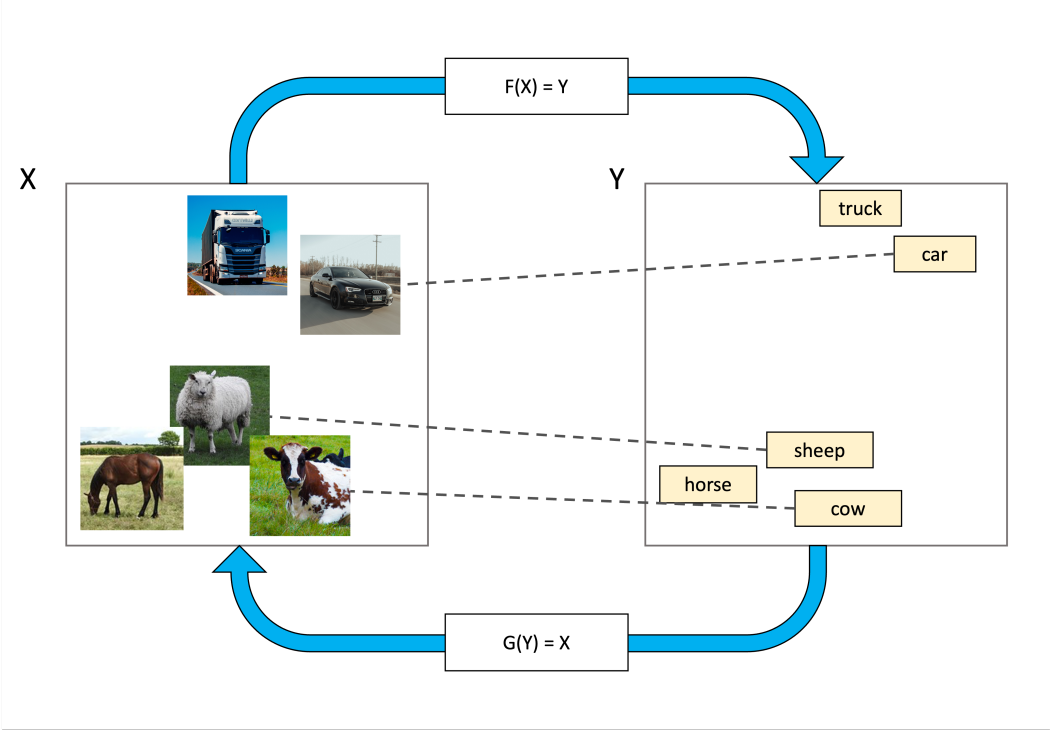


Figure 1.1: Example of systems alignment. Notice that the similarity relationships in the visual and linguistic domains mirror one another. Functions F and G learn correspondences between entire domains X and Y. Dashed lines represent known mappings for individual items. In this example, no mapping is known for 'horse' or 'truck', but the correct mapping for these items could be inferred in an unsupervised fashion based on the alignment of systems via F and G. This demonstrates how systems alignment may facilitate generalisation.

This dissertation explores the consequences of shared cross-modal structure for learning from multiple perspectives. First, it explores the role that aligning systems across naturalistic modalities could play in a relevant real-world learning problem: early concept acquisition. Second, it asks whether and how aligning shared structure between different systems facilitates efficient learning in humans. Third, it investigates whether aligning this shared structure can benefit machine learning systems attempting cross-modal learning in unsupervised or low-data environments.

1.1 Conceptual similarity relationships in mind and brain

The hypothesis that learning benefits from similarity structure depends on the assumption that humans are sensitive to similarity relationships between entities. This section explores prior work on the role of similarity relationships in humans' understanding of the world.

Similarity relationships between entities are of great importance in accounts of semantic knowledge (McRae et al., 1997), and the workings of the brain more broadly. In fact, candidate neural substrates for semantic knowledge are often evaluated using the correlations between semantic similarity judgments and the similarity in their activity patterns (Martin et al., 2018; Visser et al., 2012).

An established body of work shows evidence that cognitive and spatial relationships are handled similarly in the brain. This originates with Tolman's theory of the 'cognitive map' (Tolman, 1948), and has since found support from many studies demonstrating that spatial and conceptual relationships (across multiple modalities) are supported by shared neural mechanisms (O'keefe and Nadel, 1978; Constantinescu et al., 2016; Bao et al., 2019; Behrens et al., 2018; Bellmund et al., 2018). Theves et al. (2019) show that distances between concepts in abstract feature space are encoded in the hippocampus during concept learning. Whether this evidence is interpreted as suggesting that we navigate conceptual spaces through spatial means, or vice versa (Mok and Love, 2019), the existence of geometric structure in our neural representations of conceptual information is strongly evidenced.

Behavioural studies have demonstrated sensitivities to structural correspondences between concepts from a young age. Unger et al. (2020a) used a range of paradigms including cued recall, match verification and eye gaze, to probe the influences of taxonomic similarity (the extent to which words occur in similar contexts) and co-occurrence (the extent to which words co-occur

with each other) between concept words. They tested 4-5 year old children and adults, and found that children were consistently sensitive to co-occurrence relationships between concept nouns, and adults were consistently sensitive to both co-occurrence and taxonomic relationships.

The influence of semantic similarity relationships has also been demonstrated in the context of learning. Chen and Yu (2017) showed that semantically themed learning contexts led to improved learning across a pair of cross-situational learning experiments, and that this effect was independent of recall context. A recent series of experiments in category learning demonstrated that the perceived strength of semantic category membership distorted newly acquired image-location associations (Tompary and Thompson-Schill, 2021). These results provide evidence that learning leverages the network of existing inter-concept relationships within a semantic framework. Systems of similarity relationships are also explored within the context of analogy, discussed in more detail below.

1.1.1 Analogy and structural alignment

As humans, we are skilled at establishing correspondences between systems of structural relationships. These correspondences underpin our penchant for analogy (Gentner, 1983; Gentner and Smith, 2012), which has long been viewed as a key component of our intelligence (Holyoak, 2012; Mitchell, 2021). In many domains, analogy is deliberately employed to support learning: by mapping unfamiliar systems onto familiar ones with shared structure, we are able to efficiently integrate new information using existing knowledge frameworks (Gentner and Holyoak, 1997; Richland and Simms, 2015). Early studies demonstrated the impact of analogy on memory, showing that alignable cues yielded better recall than unalignable cues following the explicit comparison of relationships in visual scenes (Markman and Gentner, 1997).

Local structural relationships are also exploited in fast-mapping, where learning has been shown to occur based on structural correspondences be-

tween words in context - e.g., when posed with the task: 'pass the chromium tray, not the blue one', the appropriate inference about the meaning of the unknown word 'chromium' can be made (Heibeck and Markman, 1987; Carey and Bartlett, 1978). This form of perceptual alignment has also been explored as a signal in early adjective learning, and has been found to aid learning in incidental learning contexts (Shao and Gentner, 2022).

But where analogy seeks alignment between two analogs (by processes which are sometimes referred to as 'structural alignment') the work in this thesis is oriented around the possibility that entire conceptual systems could be aligned for learning, at scale. One important application of such an alignment process might be in the integration of information from multiple modalities to form unified concept representations.

1.2 Multimodal concepts in the mind and brain

As our experiences of the world and the concepts therein are fundamentally multimodal (Fernandino et al., 2016), it is no surprise that concept representations in the brain are also multimodal in nature. When we think of a cat, for example, we are able to call to mind its general appearance as well as the sounds it may make and the feel of its fur. When we think of a banana, we recall its color and form alongside its taste and how best to peel it.

The multimodal nature of concept representations in the brain has been understood for over a century (Wernicke, 1900 - see Wernicke, 1977; Kiefer and Pulvermüller, 2012). Indeed, *distributed-only* theories of semantic knowledge, which purported that distributed activations across interconnected sensory systems were the entire basis of concept representations (Patterson et al., 2007a; Barsalou et al., 2003), were once the dominant theories of semantic knowledge. While consensus on their role has evolved, distributed and experience-based neural activations remain an important part of modern theories of concept representation in the brain (Meteyard et al., 2012; Martin, 2016).

A key development in theories of semantic knowledge has been the inclusion

of a *transmodal hub*[1]: an area of the brain which stores concept knowledge independent of any specific modality. The transmodal hub was first warranted by neuropsychological evidence. Warrington (1975) found that Semantic Dementia (SD) patients consistently demonstrated selective and multimodal deficits in semantic knowledge, while their episodic memory and other cognitive functions remained intact (Warrington, 1975; Patterson et al., 2007a).

By the time of symptom onset, SD patients were found to consistently exhibit specific deterioration in the Anterior Temporal Lobes (ATL) (Patterson et al., 2007a). This was the first evidence suggesting that the region may play a role in supporting transmodal concept representations. The *hub-and-spoke* model proposed by Patterson et al. (2007a) simultaneously accounts for (a) the distributed nature of concept representations across sensory modalities and (b) the transmodal representations of concepts supported by the ATL.

Since the hub-and-spoke theory was put forward, further evidence has emerged demonstrating semantic knowledge's joint dependence on distributed multimodal representations and a transmodal hub. PET and MEG studies had previously demonstrated a role for the ATL in transmodal concept representations (review in Jefferies, 2013), and despite once uncertain evidence (Martin, 2007), corrective techniques have yielded convergence to the same result in fMRI (Visser et al., 2010, 2012; Peelen and Caramazza, 2012). Transcranial magnetic stimulation (TMS) in neurologically healthy individuals has also been used to demonstrate the role of the ATL in semantic memory (Pobric et al., 2010). Category-general deficits in object naming were temporarily induced by stimulation of the ATL, while specific deficits for manipulable objects were induced with stimulation to the motor-relevant Inferior Parietal lobe.

Computational modelling also supports the existence of a transmodal hub for semantic knowledge (Rogers et al., 2004): while linear combinations of distributed activations are insufficient to generate meaningful representations, the inclusion of a transmodal hub allows semantic models to acquire abstract representations which encode conceptual similarity as humans do (Lambon Ralph,

[1]Referred to as the *amodal hub* in some literature (Lambon Ralph, 2014)

2014). Thus, the transmodal hub's ability to capture semantic similarity relations within a high-dimensional space is a key component of arguments in its favour (Caramazza et al., 1990).

In sum, current evidence has converged on the existence of a transmodal semantic hub supported by the ATL. This transmodal hub interacts with a distributed, multimodal network to support the abstract representation of concept knowledge while maintaining links to concept-relevant sensorimotor information. This evidence demonstrates that conceptual knowledge in humans is multimodal at its core, and necessitates the integration of multimodal information. We must now review the extant knowledge on how multimodal concepts are acquired in humans, to understand the role that unsupervised alignment signals may play.

1.3 Concept learning in humans

One aim of this project is to explore the role of cross-modal alignment in human concept learning. We define a concept as a mapping between representations in multiple modalities. In this section, I describe the current state of knowledge on how multimodal representations are acquired, and describe the evidence suggesting that an unsupervised alignment-based mechanism could play a role.

Concept learning has been defined as 'tying words to evolving concept representations' (Lake and Murphy, 2021). A person's concept representations develop over the course of their lifetime, in response to their idiosyncratic experiences. A child's representation of an airplane, for example, will be far more simplistic than that of an aerospace engineer. However, the communicative role of language means that the language we attach to concepts is shared across individuals and their idiosyncratic experiences, and is robust to these idiosyncracies. Concept labels function as landmarks for coordination across individuals with varying experiences (Enfield, 2022).

Acquiring a multimodal understanding of concepts as a naive learner is a remarkable feat of cognition. Yet, it has been shown that infants can acquire an

understanding of more than 300 concepts by 16 months of age (Fenson et al., 1994). The phrase *vocabulary spurt* has been used to describe the notable increase in the rate of word acquisition in the second year of life (Bloom, 2013).

Prior work has identified a range of factors which influence how concepts are acquired. Lexical, phonological and semantic features—such as word frequency, phonological neighbourhood size, and associations with other words—have all been found to be predictive of a concept's age of acquisition (Storkel, 2009; Braginsky et al., 2016; Schneider et al., 2015; Hills et al., 2009; Stella et al., 2017). This section focuses on how different types of learning signal have been shown to contribute to concept learning process in humans. It concludes with the suggestion that there is room for unsupervised learning signals derived from cross-modal alignment to be playing a role in human concept acquisition.

The range of learning signals used by humans in acquiring correspondences between systems are outlined in this section. Throughout this discussion, parallels are drawn between machine- and human learning scenarios, and human learning is categorised according to principles perhaps more commonly discussed in machine learning: *supervised*, *semi-supervised*, *weakly-supervised*, and *unsupervised learning*. This reflects a key theme of this thesis: the value of a bidirectional interaction between cognitive science and computational modelling. This is discussed further in section 1.4.

Supervised and semi-supervised learning

The first type of learning at play is *supervised learning*, defined as learning from labelled examples. In the context of early life, this could be a caretaker pointing at a dog while saying the word 'dog'. Analogously in a machine learning context, a machine learning system learning to classify images of cats and dogs could be provided with numerous examples of images of cats and dogs, along with the correct associated labels. Based on this supervision, the

model can be trained to identify the features which distinguish images of cats from images of dogs, and to make the correct classification.

In the real world, though, even the most supervised concept-learning learning episodes - for example, pointing at an object while naming it aloud - are riddled with ambiguity. This is demonstrated by Quine's famous 'gavagai' thought experiment (Quine, 1960): if a teacher points at a rabbit hopping through a field and says 'gavagai' aloud to a naive learner, how does the learner know what 'gavagai' refers to? It could mean hopping, rabbits generally, this rabbit specifically, the rabbit's fur - the list of possibilities goes on.

Despite this ambiguity, supervised learning events demonstrably improve the development of a child's vocabulary across cultures (Shneidman and Goldin-Meadow, 2012). Constraints, such as the mutual exclusivity assumption, the taxonomic assumption and the whole-object assumption (Markman, 1990, 1994), are known to play a role resolving the ambiguity of such labelling events, perhaps rendering them 'supervised'.

However, the extent to which infants encounter supervised learning events varies greatly, both within and between cultures (Cartmill et al., 2013; Lieven, 1994), and a relatively small proportion of an infant's language exposures take the form of supervised labelling events: 60-70% of concrete nouns in child-directed speech are not in reference to the current environment or activity (Tamis-LeMonda et al., 2019; Clerkin and Smith, 2022). Learining from sporadic or infrequent labelling events can be referred to as *semi-supervised* learning.

Weakly-supervised learning

The next type of learning contributing to concept acquisition consists of learning from an imprecise supervisory signal, frequently referred to as *weakly-supervised* (Zhou, 2018). Arguably, based on the discussion above, a large number of learning events fall into this category for human learners. Adults and infants alike are capable of leveraging statistical regularities from their

environments to improve the efficiency of learning under such conditions: humans have been shown to benefit from cross-situational statistics when learning multi-modal concepts from as early as 12-months of age (Smith and Yu, 2008; Yu and Smith, 2007).

Language acquisition can also occur successfully in the near-absence of explicit instruction: infants can learn from indirect word exposures, either through overhearing or interactions not intended as learning events (Lieven, 1994; Saffran et al., 1996; Akhtar et al., 2001; Akhtar, 2005; Gampe et al., 2012; Jaswal and Markman, 2001; Shao and Gentner, 2022). In the Psychology literature, this type of learning is often referred to as *incidental learning*, but the presence of a supervisory signal in these studies—even if weak—warrants their classification as event-based learning episodes.

Unsupervised learning

Finally, we must review the evidence for *unsupervised learning* processes in cross-modal learning. While the importance of event-based learning is indisputable, some concept learning likely occurs in the absence of even weak supervision. Learning a concept involves information spanning multiple sensory inputs - for example, the concept 'bird' may include the sound of birdsong - but this information is not consistently provided in the same concept acquisition episode. Successful concept integration across multimodal systems may, therefore, benefit from asynchronous learning processes.

The finding that blind and sighted participants have similarly organised semantic activations in the brain for visual and non-visual stimuli provides evidence for such asynchronous semantic integration (Vetter et al., 2020, 2014). In developmental contexts, it has has been noted that much of an infant's linguistic exposure is not child-directed speech, and may not temporally co-occur with an associated visual referent (Lieven, 1994). It has been shown experimentally that children are capable of integrating multimodal concept information when the object and referent are presented asynchronously (Samuelson et al.,

2011).

Some research efforts have explored whether structure in naturalistic data could be used as a supervisory signal in concept learning. Machine learning systems (often described as *self-supervised* systems) do this very successfully (Harris, 1954; Pennington et al., 2014; Mikolov et al., 2013c,a; Lund and Burgess, 1996; Chen et al., 2020b; Devlin et al., 2018). It has been shown that infants are sensitive to co-occurrences in language from a young age (Unger et al., 2020b), and semantic information can be derived from the co-occurrence statistics in child-directed speech (Li et al., 2000). However, little work has been done to explore unsupervised cross-modal learning in humans.

The foundations for unsupervised cross-modal learning have been demonstrated by work showing that structural relationships are recapitulated across modalities. Early evidence demonstrated that there are redundancies in the information captured by linguistic and visual systems (Riordan and Jones, 2011), and that perceptual features can be predicted from linguistic co-occurrence data (Johns and Jones, 2012; Lewis et al., 2019). Roads and Love (2020) conducted an information analysis on unimodal embeddings across multiple modalities, which found that co-occurrence based relationships remain consistent across modalities. That is, if 'cat' and 'dog' occur in similar linguistic contexts, their corresponding referents are likely to occur in similar visual contexts. As such, it may be possible to leverage structural correspondences in learning mappings between modalities. While the semantic spaces they used were not continuous, Tompary and Thompson-Schill (2021)'s results are consistent with a form of alignment in the acqusition of new information. This alignment process could support the formation of multimodal concept representations from complex inputs. However, no scalable model of the role of alignment in learning has been proposed.

The structural consistencies across modalities identified by computational means, and our proven human sensitivity to structural relationships in mind and brain, lend plausibility to the idea that mappings between systems may

be leveraged as part of the concept acquisition process. As this section has demonstrated, this thesis explores this possibility from the joint perspectives of computational modelling and cognitive science. The following section reviews the background to this bidirectional approach.

1.4 Computational modelling and cognitive science

This section addresses the interaction of cognitive science and computational modelling in general, with a focus on learning. Both the applications of computational modelling to cognitive science and the applications of cognitive science to computational methods are discussed, to establish a precedent for the methodology used in this thesis.

1.4.1 The application of computational modelling to cognitive science

Computational modelling is widely viewed as an essential component of cognitive science (Murphy, 2011). One key reason for this is that the process of developing computational models of cognition forces theory to be formalised and specified, which in turn allows the theories mechanisms to be better understood. Along similar lines, McClelland (2009) frames the role of modelling in cognitive science as exploring 'the implications of ideas about cognitive processes'. Computational models are also valuable as a means of generating testable hypotheses (Farkaš, 2012; Stafford, 2012), and providing a process for their testing and refinement.

In other domains, computational models may be assessed on their efficiency and optimality for the task at hand. But when evaluating computational models of cognitive processes, the key questions are whether the model carries out a task as well as humans do, and whether mistakes it makes reflect the mistakes that humans make in attempting the same task. This is referred

to as a model's empirical adequacy (McClelland, 2009). It is crucial to note, though, that good model fit does not necessarily mean that the model reflects the true cognitive process. All that can be concluded is that the model is a candidate model worthy of further exploration, for example via the falsifiable hypotheses that it generates.

In this thesis, computational models are used to explore the explanatory power of cross-modal alignment in learning, and to test whether proposed mechanisms are promising candidate mechanisms of human learning behaviours.

1.4.2 The application of cognitive insights to computational models

Another benefit of using computational models within cognitive science is the ability to implement cognitive findings in machine systems. In this thesis, one aim is to use findings on how humans may learn to perform cross-modal mapping tasks via alignment, and apply this to machine learning systems, which currently struggle to achieve human performance from naturalistic stimuli.

There is a long-standing tradition of human-inspired computational systems. This is perhaps most notably exemplified by the history of connectionism, whose start in the form of Rosenblatt's perceptron as a model of storage in the brain (Rosenblatt, 1958) ultimately laid the groundwork for the parallel distributed processing at the heart of today's deep neural networks (Rumelhart et al., 1986).

The origins of deep neural nets lie in attempts to model the brain (McCulloch and Pitts, 1943) permeate many of the most successful methods and architectures used today. The field of computer vision and many of the most successful architectures developed therein is heavily based on insights about the human visual system (Hubel and Wiesel, 1959; Fukushima, 1980). While arguably modern machine learning is not principally concerned with biological plausibility, recent key advances in machine learning continue to take their in-

spiration from cognitive processes, for example the implementation of attention mechanisms to deep models (Vaswani et al., 2017; Lindsay, 2020).

The practice of implementing cognitively-inspired elements into computational systems is often motivated by a desire to address gaps between machine and human task performance. Advances in machine learning have led to machine performance surpassing human performance on a wide range of tasks - from games like chess and Go, to diagnosing cancer from medical imaging - but there are still many tasks where machine systems struggle in ways that humans do not (Lake et al., 2017).

When it comes to learning from unstructured real-world data, a popular approach is to explore developmental inputs into machine learning (Smith and Slone, 2017). As Zaadnoordijk et al. (2022) put it, infants are 'natural born intelligent systems', and thus the means by which they build their understanding of the world can be useful in teaching machine systems to extract information from the environment. Zaadnoordijk et al. (2022) suggest that infant-inspired systems could benefit from incorporating constraints which mirror the constraints on infant information processing, as well as by incorporating deliberate curriculum learning (Bengio et al., 2009) and by building in appreciations of statistical regularities across inputs from different modalities. This final suggestion is a key aim of the work presented in the current thesis, and the interplay of statistical structure and developmental trajectories is addressed in Chapter 2.

I now move to explore computational representations of conceptual spaces, with a focus on language representation, as this field of study has generated extensive advances in representational spaces. I first explore some of the key questions these approaches have raised regarding how language relates to the environment and captures meaning - questions highly relevant for this thesis, which explores mapping between language and other modalities at length.

1.5 Learning in the Chinese room

Before reviewing how concept representations are addressed in the literature, I will first set the scene by applying alignment principles to a modern adaptation of John Searle's Chinese Room thought experiment (Searle, 1980), which addresses the relationship between language and meaning. An alignment-based interpretation of this thought experiment yields an interesting new conclusion. This example demonstrates that alignment provides a novel perspective on learning, which has the potential to contribute to modern debates in computational linguistics.

The Chinese Room thought experiment serves in its original form to demonstrate that no *understanding* of language is required in order to generate a convincing linguistic output. The thought experiment in its original form is as follows: imagine an English speaker who knows no Chinese is locked in a room alone. People slip pieces of paper with Chinese characters on them under the door (the input). The English speaker has a very comprehensive manual with a series of steps, outlining how to select characters in response to the input (the program). The program guides them to produce an output, also in Chinese, which they then return under the door. Unbeknownst to the English speaker, the inputs they are receiving are questions, and their manual guides them to respond with the appropriate answers. On the other side of the door, the Chinese speakers submitting the questions could be convinced that the person in the room understands and speaks Chinese, when in fact the English speaker has no understanding at all. The intended conclusion here is that, provided the system's program is adequate to generate the appropriate response, no understanding of language is required to produce meaningful language outputs.

In the modern day, leaps forward in the performance of large language models (LLMs) have raised further questions in the mainstream about what it means to understand language. The success of LLMs is such that users come away convinced the model must understand what it is saying (Bender and

Koller, 2020). The implication that an AI agent 'understands' language seems to have been widely unsettling, and has generated a vast amount of public discourse. Of course, the word 'understanding' is loaded with implications of intelligence - perhaps even sentience. These implications are entirely unwarranted, yet LLMs' apparent ability to capture meaning, having been trained on massive volumes of text data, is impressive enough to convince people otherwise.

In an attempt to address these misconceptions, Bender and Koller (2020) argue that a connection to non-linguistic systems is a requirement for meaning. In an adaptation of the Chinese Room, Bender asks the following: if a non-Thai speaker was trapped in the National Library of Thailand, with access to all books (excluding those containing pictures or those in any other language), could this person - with infinite time - learn to understand written Thai (Bender, 2023)?

This poses a different question to the original thought experiment, in that the point of interest is not whether a system programmed to process language can be said to understand, but rather whether understanding can be achieved through language alone, independently of (or separate from) experience of the world. This is related to Harnad's symbol grounding problem (Harnad, 1990), which can be formulated as the impossibility of learning Chinese from a dictionary written in Chinese.

The symbol grounding problem and work relating to it are addressed further in the next section, but the implications of alignment in the real world are exemplified well by a proposed solution to this thought experiment.

Assuming that the individual in the Thai library had prior experience of the world, learning Thai under these circumstances could indeed be possible under an alignment account. One could align a system of similarity relationships in Thai, obtained from word usage patterns, to the known similarity relationships between entities in the world, thus learning a mapping from the entities in the world to their linguistic representations in Thai. In other words, the

lack of synchronous presentation of language and non-linguistic information is theoretically not a requirement for learning a mapping between the two.

The alignability of spaces across multiple modalities could explain, in part, how LLMs are able to do such a convincing job of emulating conceptual 'understanding', when they have only ever received text inputs. If the alignment between similarity relationships across systems was perfect, then nothing more about the meaning of individual words would be learned by 'grounding' them in the non-linguistic space via synchronous experience. One could arguably know the full extent of word meaning from words alone, assuming that they had prior experience of the world. And even without this experience, text information alone could provide an agent with similarity relationships which reflect the structure of the world in other modalities. In reality, the alignment may not be perfect, but the fact that some alignment exists means that some degree of meaning can be captured after training from linguistic input only.

The Chinese Room thought experiment and its modern descendants capture key philosophical questions around language's connection to meaning. Here, we have demonstrated that alignment perspectives could help to offer a novel interpretation of some of these problems. Moving beyond the thought experiment, we now explore prior work on these questions in more detail, to understand where alignment fits in.

1.6 Models of conceptual spaces

This section discusses computational models of conceptual spaces, and addresses debates around the value of different methods for ascertaining meaning. First, I review distributional semantic approaches, which are largely based on Frith's perspective that word meaning is derived from word context. Then I discuss arguments against this approach, which largely centre on the view that a concept's meaning depends upon establishing correspondence between language and non-linguistic systems. Finally, I argue that an alignment perspective would go some way to reconciling these views, by suggesting that

the similarity relationships within the linguistic modality facilitate correspondences to non-linguistic systems.

1.6.1 Distributional semantics

According to Harris (1954)'s distributional hypothesis, semantic similarity is dictated by the similarity of contexts in which concepts appear. This hypothesis is famously summarised by Firth (1957), as: 'you shall know a word by the company it keeps'. This is the founding principle of distributional semantic models (DSMs), which in turn form the basis of many widely employed Natural Language Processing (NLP) techniques. The following section summarises some of the key distributional semantic approaches to extracting meaning from co-occurrence statistics.

Latent Semantic Analysis (LSA) is an early example of how linguistic context has been used to extract semantic information (Landauer and Dumais, 1997). Defining the linguistic context as the words present in an entire document, often referred to as a global approach, Deerwester et al. (1990) used co-occurrence statistics across documents to eliminate dependence on the presence of individual words when identifying document topics. In LSA, words are represented as high-dimensional vectors based on their patterns of occurrence across a large set of documents. Word similarity can be measured using the cosine similarity between vectors in a reduced space.

The development of local context-based approaches to word embeddings was a further breakthrough in NLP. These approaches define 'context' as a window of fixed size C around the individual word. The Hyperspace Analogue to Language (HAL) developed by Lund and Burgess (1996) constructed co-occurrence matrices using local contexts, and demonstrated that these vectors captured some degree of a word's semantic content. Two major developments which followed were Continuous Bag-of-Words (CBOW) models and skip-gram models. CBOW models are trained to maximise the conditional probability of a target word given a set of context words (Mikolov et al., 2013a), while skip-

gram models are trained to predict the context of a given input word (Mikolov et al., 2013c,a).

GloVe word embeddings (Pennington et al., 2014) aimed to combine the benefits of global and local word-embedding approaches. They use co-occurrence statistics across text corpora to calculate ratios of co-occurrence probabilities between words. These ratios are the foundation of the learned word vectors. Transforming words into high-dimensional vectors in this way allows for vector operations to capture meaningful relationships between words. Famous examples include the operation: king - man + woman = queen (Pennington et al., 2014). Indeed, it has been shown that many analogical reasoning problems are solvable using relationships in word-embedding spaces (Peterson et al., 2020; Lu et al., 2019a; Pennington et al., 2014).

The demonstrable success of word embedding techniques (Schnabel et al., 2015) has meant that they continue to form the basis of modern NLP applications across a range of domains, from text classification (Bakshi et al., 2016; Kusner et al., 2015) and sentiment analysis (Giatsoglou et al., 2017) to information retrieval (Ganguly et al., 2015).

The extraction of semantic information from co-occurrences has also been extended beyond the linguistic domain. Sivic and Zisserman (2003) developed the Bag of Visual Words (BoV) approach to modelling scenes. This was initially used for video retrieval, but has since led to more nuanced derivation of semantic information from images. For example, Sadeghi et al. (2015) used LSA to derive representations from object co-occurrences in visual scenes, and demonstrated that these reflected taxonomic relationships between objects.

The distributional semantic hypothesis has been shown to be psychologically relevant: co-occurrence information is known to be reflected in neural activity: work from Bar (2004) and Aminoff et al. (2013) find that parahippocampal cortex activation is linked to object co-occurences. Mitchell et al. (2008) also demonstrated that embeddings are predictive of neural activity patterns. More recently, fMRI work has shown visual and text co-occurrence

statistics are predictive of responses in scene- and object-selective regions respectively (Bonner and Epstein, 2021).

Fourtassi and Dupoux (2016) find that people are able to perform zero-shot learning based on linguistic co-occurrence. Having been trained to map pseudoword labels to images (e.g 'komi' to swan), participants were exposed to sentences of pseudowords, in which the learned labels consistently co-occurred with specific novel labels (e.g 'guta'). In subsequent forced-choice tasks, participants mapped labels which had co-occurred with the learned labels onto images from the same category as their corresponding referents (e.g 'guta' was mapped onto an animal instead of a car). This demonstrates that people are sensitive to linguistic co-occurrence information in learning cross-modal mappings, bolstering the justification for the use of distributional semantics in our exploration.

1.6.2 Arguments for an embodied approach

As summarised by the Chinese Room thought experiment, it has been argued that word embeddings and other approaches based in distributional semantics cannot capture all aspects of humans' concept representations (Lake and Murphy, 2021; Bender and Koller, 2020) because human concepts are *embodied*, or rooted in perceptual experiences of the non-linguistic world (see also Barsalou 2008). Distributional approaches, meanwhile, derive the meaning of symbolic representations from their relationships to other symbolic representations, and are thus not connected to the real world. This argument can be attributed largely to Harnad (1990), who describes this as the symbol grounding problem. In this argument, learning concept representations from linguistic context is argued to be circular, like trying to learn a new language from a dictionary which is written exclusively in the language you are trying to learn.

Potential problems with ungrounded or unembodied concept representations are exemplified in Bruni et al. (2014), where it was noted that some word embeddings could not tell you that a banana was yellow, despite their

yellowness being one of the first properties humans call to mind. This comes down to the fact that text corpora do not capture basic perceptual information, despite this being fundamental to the human experience.

Having explored the symbolic and embodied approaches to understanding language, we now review efforts to reconcile these viewpoints.

1.6.3 Bridging distributional semantics and embodied approaches

It may not be helpful to place the symbolic and embodies perspectives in complete opposition to eachother. Louwerse (2008) argues that the embodied vs. symbolic language debate is outdated, and proposes the symbol interdependency hypothesis. According to this view, language comprehension can be symbolic by leveraging the interdependencies of symbols (i.e, noting the significance of the contexts in which they occur) *or* embodied, by leveraging references symbols make to their modal representations.

But perhaps more crucially, on the basis of the information presented in this Introduction, it is clear that structures within linguistic (or symbolic) spaces reflect structures in the world (Roads and Love, 2020; Riordan and Jones, 2011; Johns and Jones, 2012; Lewis et al., 2019). Because language was built on the perceptual world, the perceptual world is encoded in its statistics (Louwerse, 2018). As such, even if a representation is derived through purely symbolic means, provided that the language from which it was derived is at least partially in reference to the world, the resultant representations may be considered 'grounded' to some extent.

Further to this, if the unsupervised alignment of representations was proven possible, a stricter view of 'grounding' could be satisfied by learning correspondences to non-linguistic representations, based solely on the representations within each modality. The focus of this thesis is not explicitly to address the symbol grounding problem, but rather to understand the value of shared contextual information in learning (Vigliocco et al., 2009; Barsalou, 2008).

Having surveyed theoretical perspectives on mapping from linguistic space to meaning, and demonstrated the potential value of alignment in this context, we move to a discussion of how multimodal information has practically been applied to machine learning systems. I identify opportunities for alignment-based mechanisms to benefit these systems by providing an unsupervised signal for learning.

1.7 Multimodal machine learning

Multimodal machine learning refers to any machine learning system whose inputs span multiple modalities. In some cases, a multimodal approach is necessitated because the task of interest is multimodal in nature (Mogadala et al., 2021), for example in image captioning (see Hossain et al., 2019, for a review) and visual question answering (see Wu wt al., 2017, for review). But multimodal approaches have also been shown to improve performance in unimodal linguistic tasks such as metaphor classification (Bruni et al., 2012; Shutova et al., 2016).

Early multimodal approaches focused on building grounded embeddings, and included joint feature-topic models (Andrews et al., 2009), the concatenation of perceptual and distributional features (Johns and Jones, 2012), and the use of Canonical Correlation Analysis (CCA) to jointly project distributional and perceptual information into a lower dimensional space (Silberer and Lapata, 2012).

In recent years, multimodal language models have gained substantial traction, showing demonstrable success on a range of downstream tasks despite their training being task-agnostic (Huang et al., 2023; Chen et al., 2020d; Wang et al., 2021; Tan and Bansal, 2019; Lu et al., 2019b). Such multimodal approaches have been discussed as a promising path to something like 'artificial general intelligence'. Multimodal representations have also been shown to correspond more closely than unimodal representations to human performance in unimodal tasks (Demircan et al., 2023; Marjieh et al., 2022). These successes

reflect the importance of multimodality in our understanding of meaning.

The challenge of cross-modal translation - that is, generating mappings across modalities (Baltrušaitis et al., 2018) - has been approached using a range of multimodal techniques, across different levels of task supervision. These are reviewed below, separated into supervised and semi-supervised, weakly-supervised and unsupervised approaches.

1.7.1 Supervised and semi-supervised methods

In many cases, multimodal machine learning systems are trained on paired examples across modalities, for example by providing an image captioning system with many examples of image-caption pairs (Vinyals et al., 2015; Fang et al., 2015).

By incorporating linguistic representations of class names, image classification can be treated as a cross-modal translation task. In this formulation, generalisation to novel classes has been demonstrated by learning joint embedding spaces across visual and linguistic modalities. Lazaridou et al. (2014) compare a range of models for the supervised mapping between image and text-based distributional semantic spaces (namely a linear model, CCA, SVD and a neural network). They find that the neural network approach is most successful on zero-shot learning tasks, and yields improvements on chance in noisy, real-world datasets.

Frome et al. (2013) used supervised training to map images into a linguistic embedding space. The resultant multimodal embedding space allowed the image model to make semantically sound inferences about unseen image labels in a zero-shot learning task. A different supervised approach from Socher et al. (2013) trained a linear mapping from image space to linguistic space, where an outlier detector chose either to (a) map it into an existing linguistic category using a classifier, or (b) map it onto one of two 'outlier' classes in the multimodal space. While successful, this was only tested for 8 trained image categories and 2 outlier categories. Akata et al. (2015) scale this up,

learning a mapping from text-embedding space to image-embedding space for the purpose of fine-grained image classification. By training on a set of known classifications, their models are able to generalise classification to test examples in unseen classes.

Despite their being trained on some known concepts, the application of these models to zero-shot learning demonstrates the utility of continuous mappings between semantic spaces for generalisation tasks.

1.7.2 Weakly-supervised methods

A weakly supervised example of the alignment of modalities is found in Sigurdsson et al. (2020), where weak supervision for a translation task is provided by pairing monolingual instructional videos with their associated audio tracks. The videos serve to ground the unimodal audio tracks in a shared visual space. The supervision is rendered 'weak' by the loose association of the audio and the video, and the fact that the videos are non-identical across languages.

Lazaridou et al. (2016) successfully train models to pair words and objects from noisy temporally co-occurring verbal and visual inputs alongside social cues. This is achieved by using child-directed utterances to predict (i) the next word in the utterance and (ii) the objects present in the co-occurring visual input.

In a weakly-supervised image captioning task, success has been demonstrated by projecting images and captions into a latent space, which is aligned using the concepts common to captioning sentences and images (Laina et al., 2019).

1.7.3 Unsupervised methods

Unsupervised methods have not generally been used to find mappings between entities across modalities, and are instead common in cross-modal generative translations tasks, where the output is more open-ended. Generative adversarial networks (GANs) have been used for unsupervised image caption generation

(Shetty et al., 2017; Dai et al., 2017; Gu et al., 2019; Feng et al., 2019), and for text-to-image synthesis (Reed et al., 2016).

Optimal transport methods have been applied to structural alignment problems framed as graph matching (Titouan et al., 2019; Seguy et al., 2017). These methods were built to find mappings between distributions, and to optimise the mapping by minimising the cost of transporting one distribution to another distribution. Adaptations of these methods have been successfully applied to image-text domain mapping problems, such as visual-question answering (Chen et al., 2020a).

In this Chapter 4 of thesis, these methods are adapted and tested on unsupervised cross-modal alignment problems, in an effort to enable machines to learn cross-modal mappings from environmental signals. While these methods are utilised by prior work in multimodal machine learning, there is also inspiration to be drawn from other subfields of ML, explored in the following section.

1.8 Alternative approaches to alignment

Finally, looking beyond multimodal learning, it is valuable to explore methods which have been used for unsupervised and semi-supervised alignment tasks in other domains.

An early demonstration of alignment based on similarity structure was given in Goldstone and Rogosky (2002), in an effort to demonstrate that individuals' concept representations could be aligned, even in the face of noise across systems. The method employed here was a constraint satisfaction network, which placed items in correspondence based on the similarity of their similarity structures.

Mikolov et al. (2013b) applied a semi-supervised alignment approach for the task of machine translation. They use a relatively small number of known correspondences between languages to align distributions of monolingual word embeddings. The success of this technique is grounded in the authors' obser-

vation that different monolingual embeddings share a common distributional character, much as we have argued is the case for linguistic and non-linguistic systems.

On the basis of Mikolov et al. (2013b)'s findings, the foundations for many unsupervised alignment techniques originate in machine translation. Zhang et al. (2017) used adversarial training followed by the Earth Mover's distance, and Conneau et al. (2017) similarly use adversarial training to find a translation matrix W, but attain superior performance by refining their model using Procrustes method (Artetxe et al., 2016) followed by a similarity metric called Cross-domain Local Similarity Scaling (CSLS). CSLS helps to create a one-to-one mapping of source domain points to target domain points, by penalising matches to points with high degrees of 'hubness' (i.e, points which have many nearest neighbours). Conneau and colleagues' unsupervised translation technique performs as well as supervised machine translation baselines for some language pairs. Similar methods have been applied in conjunction with RNNs to perform translation between programming language sequences (Lachaux et al., 2020).

In the visual domain, unsupervised (or self-supervised) image translation has been very successful, with *CycleGAN* being a pioneering approach here (Zhu et al., 2017). In the absence of paired examples for image-to-image translation, this method uses the generative adversarial loss alongside a cycle-consistency loss, to simultaneously ensure successful translation and retention of image contents. The cycle-consistency loss is a form of self-supervision, as the loss term is calculated between the original image and its reconstruction in the original space once it has passed via another space.

For linear transformations, Richardson and Weiss (2021) showed that a simple algorithm which learned linear mappings between domains via the Iterative Closest Points (ICP) algortihm, was able to achieve remarkable success in many unsupervised image-to-image translation tasks. Indeed, it outperformed CycleGAN in linear tasks such as the rotation and vertical flipping of images,

and performed well in more complex tasks such as inpainting and colorisation.

An alternative framing of the alignment problem highlights another candidate unsupervised algorithm. Viewing cross-modal mapping as a one-player game with a high branching factor, Monte Carlo Tree Search (MCTS) emerges as a candidate for searching for the best mapping (Browne et al., 2012). MCTS has been successfully applied to problems with very high degrees of branching, with a particularly famous example being the champion-beating algorithm which learned to play the board game Go (Silver et al., 2016). The approach has since been generalised for extension to other games (Silver et al., 2018), and utilises deep learning models to estimate state values while conducting MCTS. Pinheiro et al. (2016) successfully performed graph-matching using MCTS formulated as a single-player game as a means of evaluating node pairings.

Along with the approaches from the previous section, these methods are adapted for the pursuit of machine learning applications of alignment in this thesis.

1.9 Overview of this thesis

Building on the literature reviewed here, this thesis investigates systems alignment from multiple perspectives, developing a new understanding of the role these event-agnostic signals could play in cross-system learning, for humans and machines. First, systems alignment is explored as a beneficial signal in early concept learning. Next, a behavioural study tests whether humans utilise alignment signals in learning when other learning signals are present. Finally, building on models of this alignment-based learning in humans, the candidate machine learning methods discussed in this literature review are applied as a means of promoting human-like ML in an unsupervised cross-modal learning context.

In Chapter 2, the possibility that alignment contributes to early concept learning is explored through a series of simulation studies. As discussed in this

literature review, it is established that supervisory signals for concept learning are infrequent and ambiguous in children's early lives. There are key outstanding questions around how children use the structure of their environments to learn in the absence of supervision, or to resolve the ambiguity in weak supervisory signals. In this chapter, it is found that children's early concepts are close to optimal for inferring novel concepts through systems alignment, enabling agents to correctly infer more than 85% of visual-word mappings without supervision. A structural analysis of the knowledge states that facilitated successful alignment found that they wee distinguished by their dense similarity relationships, both within the knowledge state and with respect to the as-yet unknown concepts. Artificial agents using these distinguishing structural features to select concepts proved highly effective, both in environments mirroring children's conceptual world and those that exclude the concepts that children commonly acquire. This shows that machine learning systems serve to benefit from insights into how humans learn from naturalistic inputs.

In light of Roads and Love (2020)'s finding that systems are alignable across modalities, and the findings of Chapter 2 which show that children's early concepts are strong facilitators of learning by alignment, the next question addressed within this thesis is: do humans learn from alignment signals when they are available? In a paired associate learning task presented in Chapter 3, I found that learning was more efficient and more successful when the systems across which pairs were learned had an underlying alignable structure. This suggests that, in the real world, humans could indeed benefit from alignable information across modalities when learning multimodal representations. Furthermore, participants who learned to map between alignable systems were able to generalise successfully to a completely novel pair of stimuli, performing successful zero-shot learning. When models were fitted to participant behaviour, it was found that a model with an asynchronous alignment mechanism provided the best account of how participants learned - not only when the underlying systems were aligned, but also when they were not. This

suggests that humans may apply alignment processes to learning problems by default, even when it is not beneficial for the task at hand.

Finally, in Chapter 4, I explore applications of alignment to machine learning efforts in cross-modal learning. First, I assess a series of modifications to alignment scoring metrics insipred by psychological literature and the findings from the investigation of early concepts, to maximise the chances of algorithm success. Then, I test a slate of algorithms drawn from a range of machine-learning subfields on the problem of unsupervised cross-modal alignment at different scales. Finally, I examine the potential for alignment to improve the performance on the cross-modal ML task of image classification, by functioning as a prior across unseen classes.

Chapter 2

Alignment in children's early concepts

2.1 Introduction

Recent work, reviewed in the previous chapter, has demonstrated that valuable signals exist which may facilitate learning via unsupervised and asynchronous mechanisms in the real world: similarity relationships between concepts have been shown to be upheld across multiple modalities. For example, concepts which are discussed in similar contexts (such as 'car' and 'truck') are likely to also appear in similar visual contexts. Based on this signal, learning could proceed in an unsupervised fashion by identifying structural idiosyncrasies that are present in both modalities and then mapping between modalities in an asynchronous process of systems alignment.

Given the challenge that concept learning presents for a naive learner (Quine, 1960), children would be strong candidates for taking advantage of alignment signals in learning. In this chapter, evidence suggesting a role for systems alignment in early concept learning is presented. First, we demonstrate the utility of alignment as a means of inferring cross-modal mappings. We then show that learning by alignment is preferentially supported by concepts acquired early in life, suggesting that early-acquired concepts may be privileged with respect to systems alignment. We then identify common struc-

tural features of early-acquired knowledge states in high-dimensional semantic spaces, and use generative modelling to demonstrate that these can be exploited to produce knowledge states which are optimally positioned for learning by alignment. Our findings contribute to the existing literature which suggests a role for these event-independent, alignment-based learning processes in acquiring cross-modal representations of the world.

While modes of learning are diverse, research predominantly focuses on *event-based* learning. Event-based learning includes popular forms of supervised, semi-supervised, weakly supervised, and unsupervised learning. A supervised learning event occurs when, for example, a child's caregiver points to a dog and labels it as a "dog". A weakly- or semi-supervised event may occur when a child overhears a conversation between two adults. Event-based learning is unquestionably an effective route for human learning, but we argue that people also use an additional, underappreciated mode of learning that is distinct from event-based learning.

Consider how remarkable it is that children can cut through noisy labels and learn from weakly supervised events. Even the most direct labelling event is ambiguous in the real world, as labelling events are heavily underconstrained (Quine, 1960; Markman, 1990, 1994). Yet infants can learn from indirect word exposure, either through overhearing or interactions not intended as learning events (Saffran et al., 1996; Akhtar et al., 2001; Akhtar, 2005; Gampe et al., 2012; Jaswal and Markman, 2001; Shao and Gentner, 2022). They can also resolve ambiguous labels by combining information across different events, i.e. cross-situational statistics (Yu and Smith, 2007). Like self-supervised machine learning systems that use structure in the data as a supervisory signal (Harris, 1954; Pennington et al., 2014; Mikolov et al., 2013c,a; Lund and Burgess, 1996; Chen et al., 2020b; Devlin et al., 2018), children may use co-occurrence information to infer meaning from natural language. Infants are sensitive to co-occurrences in language from a young age (Unger et al., 2020b), and semantic information can be derived from the co-occurrence statistics in child-directed

speech (Li et al., 2000). All of these learning feats support the idea that children have a profound ability to infer conceptual relationships, even when those relationships are not directly observed.

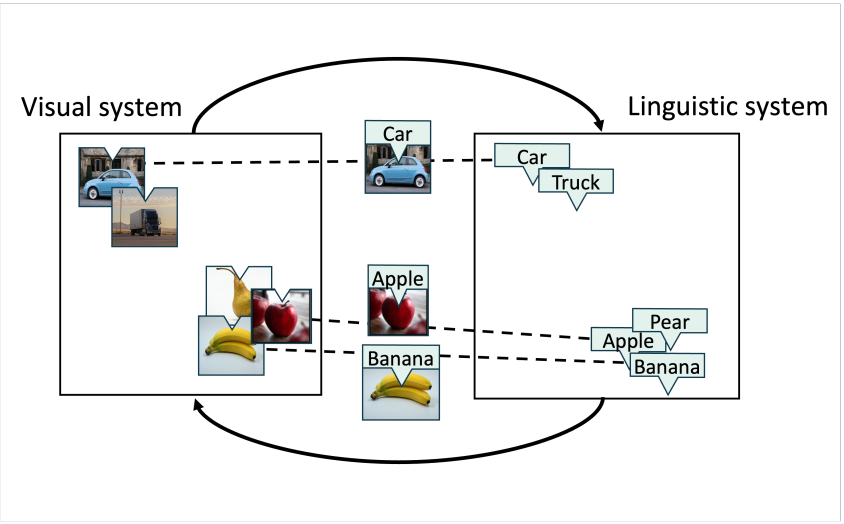


Figure 2.1: Visualisation of systems alignment in concept learning. Dashed lines between systems represent known word-object mappings. Here, the agent in question knows the mapping between {"apple", "banana", "car"} and the relevant visual objects, but does not know the mappings for the words {"pear", "truck"}. Based on the similarity relationships within the systems, however, the agent could make an accurate inference about which item was "pear" and which was "truck" when presented with the two objects. Having heard the words truck and pear being used in context, they would know that "truck co-occurs with words like "road" and "drive", while "pear" co-occurs with words like "eat" and "yummy". Similarly, having seen trucks and pears in the world, they would know that the visual context of a truck is likely to be outdoors on the highway (much like a car), while a pear is more likely to be found in a kitchen or fruit bowl (much like an apple or a banana). Therefore, if presented with the two objects and asked 'Which is the *truck* and which is the *pear*?', they could make the appropriate *forced choice*

In this work, we present evidence that children can exploit information that transcends individual events to align entire systems (e.g., to discover a mapping between visual and a word space), which we refer to as systems alignment. To re-iterate, we define systems alignment as the use of idiosyncratic similarity relations that are mirrored across multiple systems to perform a cross-system mapping (Figure 2.1).

As stated previously, we define a concept as a correct mapping across systems (e.g., the correct mapping between the word 'car' and the corresponding visual object). We refer to the set of known concepts as the *knowledge state*. In practice, prior knowledge of some concepts, such as knowing the label "car" maps to an image of a car, facilitates or bootstraps systems alignment (Figure 2.1). According to systems alignment, the more that is known, the easier it becomes to infer new knowledge. For example, based on knowing the mapping for car, we predict that a child could infer the name for a truck without ever

experiencing the verbal label "truck" co-occurring with the visual experience of a truck (Figure 2.1). Systems alignment may help explain why children's vocabularies rapidly expand after around 50 words are known (Bloom, 2013; McCarthy, 1946).

Unlike event-based learning which relies on temporally proximate information, systems alignment can be asynchronous such that information is acquired at different times in the visual and linguistic systems and can be aligned at some later time absent either input. This is a key distinction between this mechanism and previous multimodal learning approaches (Roy and Pentland, 2002; Huang et al., 2023). The asynchronous nature of alignment may help explain how label-referent mappings are learned despite their relatively infrequent co-occurrence in children's sensory input: recordings obtained from cameras mounted on children's heads in naturalistic environments reveal that the simultaneous experience of a visual object and its corresponding label is rare, with absent objects frequently being referenced, and visual objects not being named (Clerkin and Smith, 2022). Further, 60-70% of concrete nouns in child-directed speech are not in reference to the current environment or activity (Tamis-LeMonda et al., 2019).

An example of this is shown in Figure 2.2: based on a documentary voiceover heard on a prior day, a child at a zoo could use alignment to map a previously unseen animal to an animal name she has heard before. Alignment could also facilitate learning asynchronously via known processes of memory replay (Barry and Love, 2023).

One key question is whether the information present in our natural environment can support systems alignment. Roads and Love (2020) answered this question in the affirmative, demonstrating that when systems—derived from environmental measures—were aligned, their mirrored similarity structure—which they referred to as an *alignment score*—was higher than other (incorrect) mappings between the systems. Thus, in principle, an algorithm that maximised alignment score could achieve systems alignment. Here, we address a

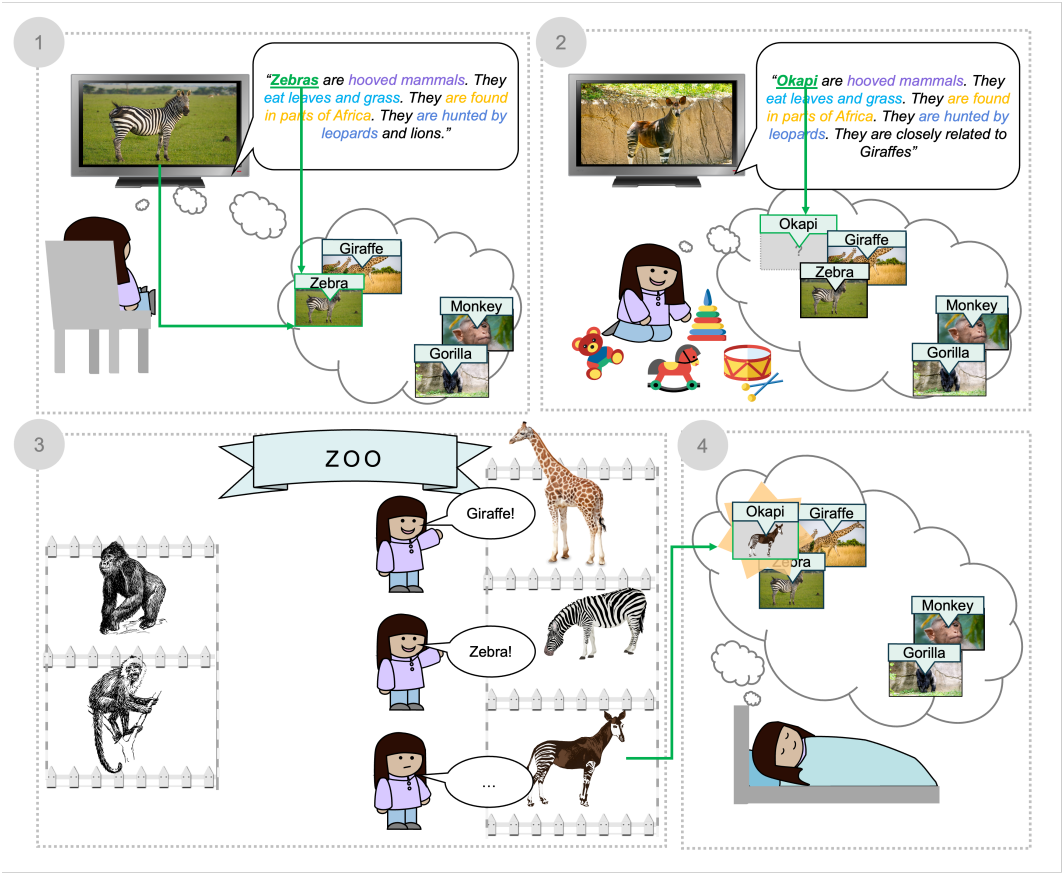


Figure 2.2: Illustration of asynchronous learning through everyday experiences. Thought bubbles depict the child's knowledge state, with visual and linguistic systems overlaid. (1) The child watches a nature documentary, where she learns about zebras from synchronous visual and linguistic input. Zebra is added to her knowledge state by this event-based learning process. (2) She begins playing with toys with her back to the television. While she is no longer watching the TV, she can still hear the documentary audio describing okapi. The descriptions of okapi and zebras are very similar, which leads to 'Okapi' being positioned in linguistic space close to 'Zebra'. Note that she does not have to understand the meaning of all words surrounding 'Zebra' and 'Okapi' for this similarity relation to be acquired. (3) Later, the child visits the zoo. From previous experiences, she can label the giraffe and the zebra. She sees an unknown animal in a nearby enclosure, which shares visual similarities with the the giraffe and the zebra. (4) Using the asynchronous inputs in different modalities, she is able to infer that the unknown animal at the zoo is likely an 'Okapi'. This is possible via alignment of visual and linguistic systems.

second question, namely could children use systems alignment to learn the meaning of words? Chapter 3 addresses another key question, testing empirically whether people engage in systems alignment when learning, and finding that they do (Aho et al., 2022).

Our consideration of systems alignment in a developmental context is novel with respect to prior work. Relevant prior work on fast-mapping demonstrates alignment effects on a local scale (e.g. 'pass the chromium tray, not the blue one', where 'chromium' is a previously unknown label) (Carey and Bartlett, 1978; Heibeck and Markman, 1987). Perceptual alignment has been explored as a signal in early adjective learning, and has been found to aid learning in incidental learning contexts (Shao and Gentner, 2022). Analogies between

word forms may help children learn to read (Goswami, 1986). But where prior work on analogy (Gentner, 1983) and alignment processes (Liu and Lupyan, 2023) has been restricted to local contexts, we argue that systems alignment could be performed between entire systems of relationships, such as across modalities to promote cross-modal learning.

Besides alignment, prior work has identified a range of factors which influence how concepts are acquired. Constraints, such as the mutual exclusivity assumption, the taxonomic assumption and the whole-object assumption (Markman, 1990, 1994), are known to play a role in ambiguous labelling events. Lexical, phonological and semantic features—such as word frequency, phonological neighbourhood size, and associations with other words—have all been found to be predictive of a concept's age of acquisition (Storkel, 2009; Braginsky et al., 2016; Schneider et al., 2015). Structural analyses of semantic networks have also identified patterns in how conceptual knowledge develops in early life (Hills et al., 2009; Stella et al., 2017; Steyvers and Tenenbaum, 2005), but the influence of structural factors in unsupervised cross-modal learning has not yet been explored. Here, we consider whether systems alignment can explain aspects of how children acquire word meanings in a manner that complements existing explanations.

We take a systems alignment view, solely concerning ourselves with factors related to the structure of similarity relationships between concepts, within a system. If children engage in systems alignment, then concepts that readily align across systems should be preferentially acquired (Roads and Love, 2020), forming a basis for subsequent learning. To foreshadow our results, artificial learning agents that are seeded with concepts acquired early by children better assimilate new conceptual knowledge through systems alignment. We proceed to investigate whether there are quantifiable structural underpinnings of this alignment effect within semantic spaces. What is it about early-acquired concepts and their relationships that allows for new conceptual knowledge to be more readily aligned? Our view predicts that knowledge states which yield dis-

tinctive similarity relationships for unknown concepts will be preferred in early life. In line with this prediction, structural analysis reveals distinctive characteristics of the similarity relationships of early-acquired knowledge states. To assess the generalisability of these structural features for improving alignment performance, we train generative agents to build knowledge states by optimising these structural parameters. Consistent with our alignment-based view, we find that agents that build their knowledge states based on these structural features outperform all other agents in their ability to learn by alignment.

2.2 Materials

2.2.1 Image embeddings

The image embeddings used here are those used in Roads and Love (2020), derived by applying the GloVe algorithm (Pennington et al., 2014) to the Open Images V4 dataset (boxes subset) (Kuznetsova et al., 2020). Open Images V4 is comprised of approximately 9.2 million images, all annotated to identify which of over 19,000 object classes they contain. Roads and Love (2020) construct a co-occurrence matrix by counting the images in which each object class co-occurs with each other class. This matrix is inputted to the GloVe algorithm, which generates the 10-dimensional image embeddings we use.

2.2.2 Word embeddings

We compared large-scale pre-trained word embeddings to word embeddings derived from child-directed speech, to choose the most suitable for this study.

Pre-trained word embeddings

The pre-trained word embeddings were 50-dimensional GloVe text embeddings (Pennington et al., 2014). These embeddings are trained on 6 billion tokens from the Wikipedia2014 + GigaWord5 text corpus. The resultant vocabulary size is 400,000 tokens.

Word embeddings from child-directed speech

It is important to consider the use of word embeddings from child-directed speech as possible relevant models of linguistic space for this study.

We inferred embeddings from the North American English subset of the CHILDES database (MacWhinney, 2000), which is comprised of transcripts of conversations and interactions with children compiled across 49 different studies. Each transcript in the database was treated as a document by the GloVe algorithm. After pre-processing to extract child-directed speech inputs and remove punctuation, the compiled corpus was inputted into the GloVe algorithm. The resultant corpus contained 4 million tokens, and had a vocabulary size of 12,252. The algorithm was run with a output vector size of 50 and a window size of 10. The algorithm ran for 1,000 iterations. The minimum count of word occurrences in order for a word to be included in the GloVe algorithm was 5.

Choice of word embeddings

When selecting the appropriate embeddings for the task, it was crucial to note that systems alignment is driven by similarity relationships between concepts, not knowledge of the concepts themselves. Preserving similarity relationships across systems is all that is required for systems alignment. Thus, a child may have different knowledge of a *car*, a *truck*, and a *shark* than an adult (Hills, 2013). However, like an adult, the child may still judge the *car* as more similar to the *truck* than the *shark*.

Analyses of the resultant embedding spaces find that this is true. For the concrete concepts explored in this study (i.e, the concepts which exist in both the word and image embeddings), key similarity relationships are preserved for child-directed speech embeddings and pretrained word embeddings. The example described above is shown in Figure 2.3. In this Figure, embeddings for the items *{car, truck, whale, shark}* are projected into 2D space using Principal Component Analysis (PCA). The results demonstrate that the similarity

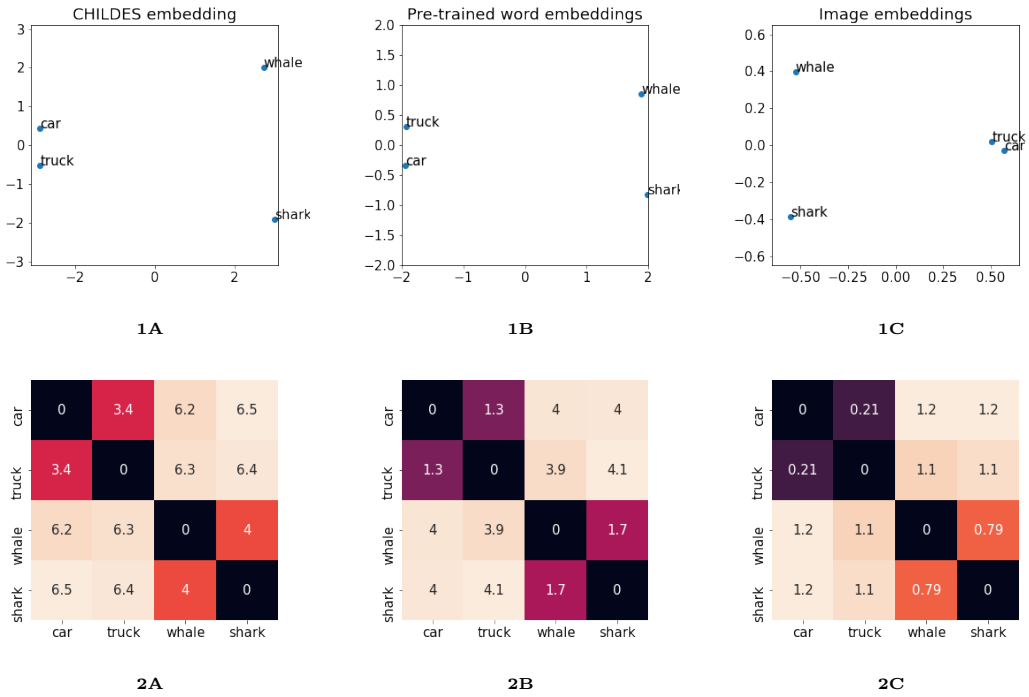


Figure 2.3: Demonstration of the consistency of similarity relations across child-directed embeddings, pre-trained GloVe embeddings and image embeddings. **1A-C**: 2D projections of the embeddings for the items {*car, truck, whale, shark*} in each embedding space. All embeddings are projected into 2D using principal component analysis (PCA). PCA is conducted on this set of four concepts in each space. **2A-C**: Pairwise distance matrices for the four example concepts. As stated in the text, the similarity relationships within the different embedding spaces are strongly consistent: in all cases, it is true that a truck is more similar to a car than it is to a shark or a whale, as the pairwise distance for the dyad car-truck is substantially lower than the pairwise distance for car-shark or car-whale. The pairwise Pearson correlations between distance matrices for this set of items are as follows: $\rho_{\text{childes-pretrained}} = 1.00, \rho_{\text{childes-image}} = 0.93, \rho_{\text{pretrained-image}} = 0.93$.

relationships for this set of items are recapitulated consistently across child-directed embeddings, pre-trained GloVe embeddings and image embeddings: in all three systems, *car* is indeed much closer to a *truck* than it is to *shark*.

Another example is provided in Figure 2.4, which demonstrates that similarity relationships which capture specific elements of meaning are also preserved across the three embedding spaces. In this demonstration, the embeddings for the items {*car, truck, airplane, cat, dog, bird*} are shown, again projected into 2D using PCA. Here, the notable result is that all three embedding systems place the items roughly on the vertices of a quadrilateral, where items can be split appropriately by two classification vectors: animal/vehicle, and flightless/flying.

This demonstrates that meaningful relationships are captured by child-directed and pre-trained word embeddings in similar ways. However, the corpus used to train child-directed speech embeddings is substantially smaller than that used for the pre-trained embeddings. It has been shown by prior

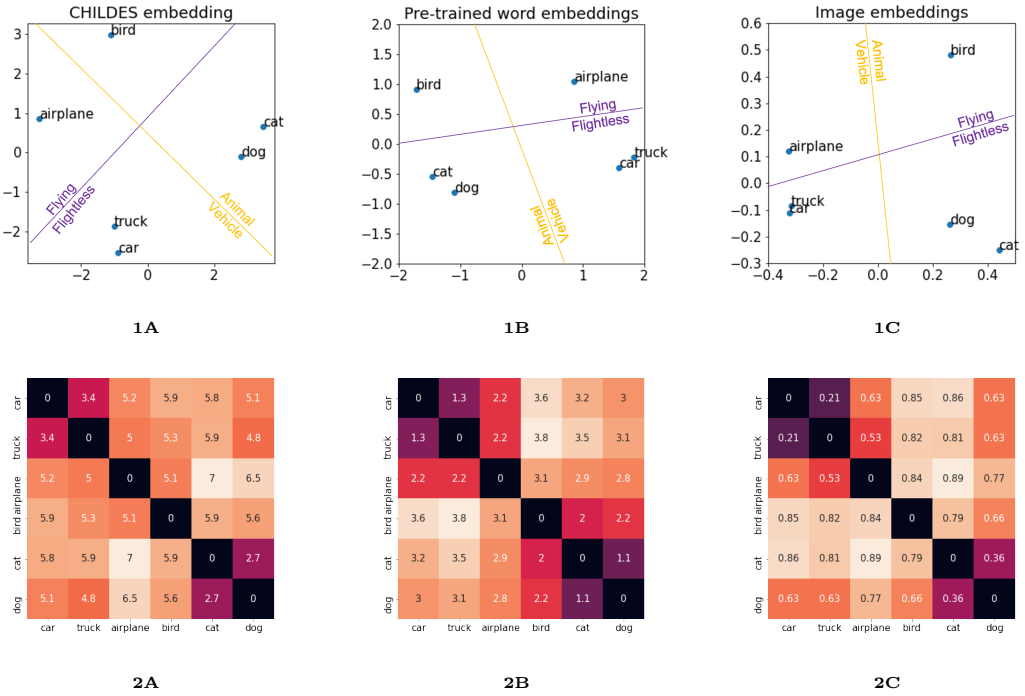


Figure 2.4: Demonstration of the consistency and meaning in similarity relations across child-directed embeddings, pre-trained GloVe embeddings and image embeddings. **1A-C**: 2D projections of the embeddings for the items {*car, truck, airplane, bird, cat, dog*} in each embedding space. All embeddings are projected into 2D using PCA performed on the item set in each space individually. Purple and orange lines demonstrate that, for all embeddings, the items can be appropriately classified on two highly salient dimensions: Animal/Vehicle and Flightless/Flying. **2A-C**: Pairwise distance matrices for the example concepts. The pairwise Pearson correlations between distance matrices for this set of items are as follows: $\rho_{\text{childes-pretrained}} = 0.63, \rho_{\text{childes-image}} = 0.86, \rho_{\text{pretrained-image}} = 0.81$.

work that corpora of comparable size to CHILDES consistently yield unstable embeddings (Antoniak and Mimno, 2018). Thus, when we expand our scope and look at larger systems of items, it may not be safe to assume that a dataset of this size will yield representatively stable embeddings.

To assess the impact corpus size may be having on the CHILDES embeddings, and to assess the extent to which CHILDES embeddings and pre-trained embeddings yield comparable similarity relationships once corpus size is accounted for, subsets of the training data used for the pre-trained GloVe embeddings were used. These subsets came from the enwik8[1] dataset. The enwik8 corpus consists of a Wikipedia dump, which is also found in the training data for original pre-trained GloVe embeddings. For our purposes, enwik8 was pre-processed such that each Wikipedia article was represented as a distinct document. 20 CHILDES-sized sample corpora were randomly sampled from the enwik8 dataset, and an embedding was inferred for each using the

[1]Obtained from http://mattmahoney.net/dc/textdata.html

GloVe algorithm, with same parameters as were used to infer the CHILDES embeddings.

A noise estimate for CHILDES-sized corpora was obtained by evaluating the similarity between each of these 20 sampled enwik8 embeddings and the original GloVe embedding. The results for the downsampled enwik8 corpora demonstrate the challenge of comparing the embeddings derived from corpora of different sizes: even when the corpus is included in the training set of a larger embedding, as is the case for samples from the enwik8 corpora and the pre-trained GloVe embeddings, the similarity scores are relatively low (see Figure 2.5). Against this backdrop, the child-directed embeddings exhibit the similarity performance one would expect given their small size relative to the original GloVe embeddings.

Given these results, pre-trained word embeddings were selected as the primary text embeddings for the analyses in this study, owing to (a) the large size of the training corpus, which has been shown to significantly impact embedding stability (Antoniak and Mimno, 2018), (b) their established correspondence to human semantic judgments of language (Pereira et al., 2016) and (c) the finding that child-directed speech embeddings correlate as highly with these embeddings as these embeddings do with themselves, presented below. With point (c) in mind, together with the findings that key similarity relationships are preserved in child-directed speech embeddings as demonstrated above, it is likely that alignment-based findings using pre-trained embeddings will also be relevant for child-directed speech, however points (a) and (b) mean that the data is currently not sufficient to rely on child-directed speech embeddings as representative of children's semantic spaces.

To bolster the relevance of our findings for early concept learning in the real world, the first experiment presented below was also conducted using the CHILDES embeddings, and the key result was replicated.

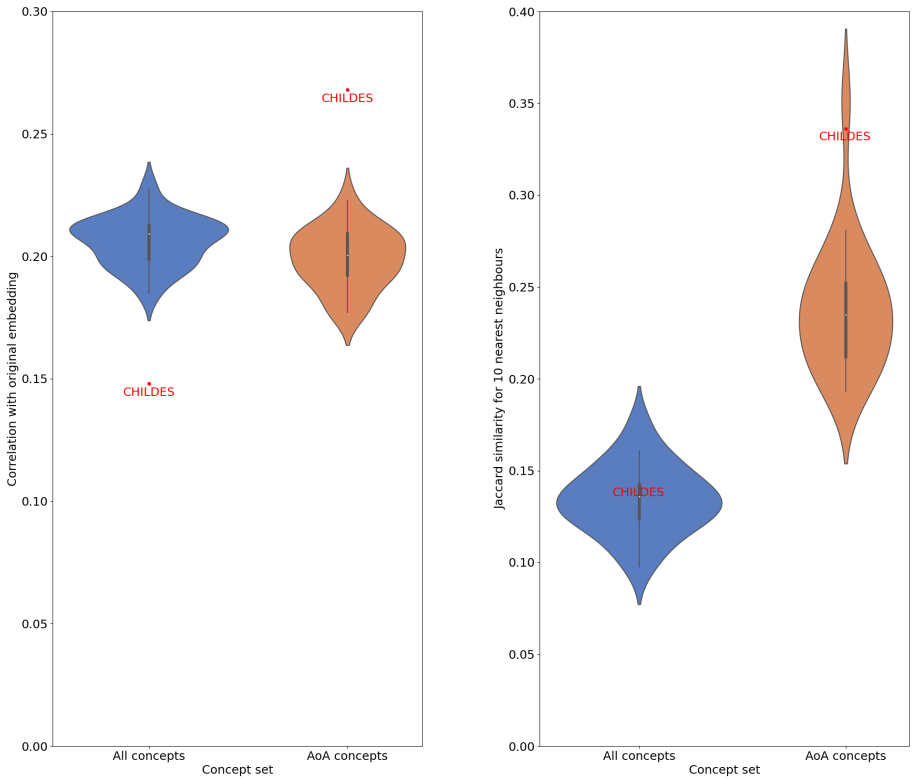


Figure 2.5: [Correspondence of CHILDES embeddings and embeddings inferred from the downsampled enwik8 corpus with pre-trained GloVe embeddings. (a) Correlations of pairwise relationships and (b) Jaccard similarity scores for 10-nearest-neighbours of concepts across 20 embeddings of comparable size to the CHILDES corpus, sampled from the enwik8 corpus. The corresponding performances of CHILDES on the relevant metrics are also shown. Performance is shown for all concepts in the set (blue) and for early-acquired concepts only (orange). On both similarity measures, the CHILDES embeddings perform comparably to the comparably-sized samples from the GloVe embedding training corpus. CHILDES embeddings even outperform the enwik8 embeddings of comparable size for early-acquired words. This shows that the CHILDES embeddings are as similar to the large-scale pre-trained GloVe emebeddings as they could be expected to be, given their corpus size. Thus, there is no evidence here to suggest that the similarity relationships for children are substantially different to those for adults, for the concepts used in this contribution.

2.2.3 Age-of-acquisition data

Age-of-acquisition (AoA) data taken from Frank et al. (2017)'s WordBank dataset. WordBank aggregates experimental results using MacArthur-Bates Communicative Development Inventories (MB-CDI) (Fenson et al., 2007). We used the English (American) dataset, containing data from linguistic development trajectories of 8,300 children. Specifically, we used the item trajectory dataset, which reported the proportion of children who could produce each word by each month of age. This data is obtained from parental reports of children's word production. This dataset contains monthwise probabilities of the acquisition of 680 words overall. We pre-process the dataset by taking the

subset of WordBank words which exist in the intersection of our word and image embeddings. There are 418 words in the word/image intersection, and of these 138 words are present in the WordBank dataset. These 138 words comprise the AoA concept set, and the full set of 418 concepts in the image/word embedding intersection intersection comprise the control concept set.

AoA data is available for children from 16 to 30 months of age. However, as the MB-CDI is an index of representative words for early vocabulary, and not a comprehensive review of a child's entire vocabulary, it is known that MB-CDI results diverge from true vocabulary size as MB-CDI scores increase (and, typically, as a child gets older) (Fenson et al., 1994, 2007). This is because as the vocabulary expands, the representative words which comprise the MB-CDI become less likely to capture the idiosyncrasies of an individual child's vocabulary. Mayor & Plunkett modelled the extent of the divergence (Mayor and Plunkett, 2011), and provided estimates for the proportion of the vocabulary which is not captured within the MB-CDI at each month of age.

For our probabilistic interpretation of the MB-CDI data, we require the assumption that the probability of a child having acquired any concept outside of the index is approximately 0. Therefore, all of our modelling and analyses are performed using WordBank data for months 16-24 only, where the index is likely to capture close to 100% of a child's vocabulary.

We assume that the order in which words are produced corresponds to the order in which words are 'known'. For our purposes, a word is 'known' when a correspondence is established to its visual form from its linguistic form (i.e, an agent can correctly label a picture of a dog as a 'dog'). Estimating children's knowledge bases using production norms likely introduces noise into our analyses and may underestimate semantic knowledge because other factors, including phonological, will influence which words children produce.

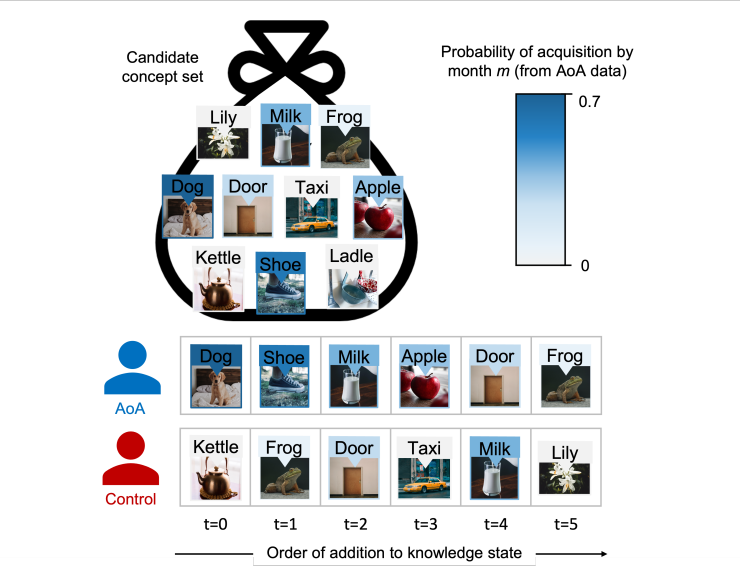


Figure 2.6: An illustrative example of how knowledge states expand in simulated agents. In this example, six concepts ($n_m = 6$) are added to each agent's knowledge state. The AoA agent's knowledge state grows in accordance with the probabilities of each concept's acquisition by month m in the AoA data (i.e., the agent acquires concepts typical of children). The Control agent ignores AoA information – the concepts added to its knowledge state are randomly drawn from the full set of concepts (see Materials and Methods).

2.3 Forced choice by alignment

Our first set of simulations tested the alignment capabilities of early-acquired concepts by examining the ease with which new concepts could be learned without instruction, based solely on relationships with known concepts.

2.3.1 Methods

Knowledge trajectory simulation

We compare the efficacy of knowledge assimilation for two *agent conditions*: agents whose knowledge states are based on real Age-of-Acquisition data (Frank et al., 2017) (AoA agent condition), and agents whose knowledge states are randomly selected (Control agent condition). Idealised examples of the knowledge state simulation process for each agent are shown in Figure 2.6.

To simulate knowledge trajectories, we calculated the mean number of concepts n_m acquired in each month m of the Age-of-Acquisition data. This is achieved by summing the probabilities of acquisition across all concepts in each month, $n_m = \sum_{i=0}^{N} p_{i,m}$, where N is the total number of concepts in our AoA set, and rounding to the nearest integer. This produces the sequence

given in Table 2.1. At each month for each agent type, we therefore have a simulated *knowledge state* which contains N_m concepts, where $N_m = \sum_{j=16}^{m} n_j$. We generate sequences of acquired concepts under two conditions:

- **AoA:** New items are selected from a probability distribution across items in the WordBank dataset (see *Materials*). The probability distribution is generated by normalising the probabilities of acquisition for all concepts in the WordBank inventory which have not yet been added to the simulated sequence, such that the probabilties sum to 1.
- **Control:** New items are randomly selected from all items in the intersection of word and image embeddings, which have not yet been added to the simulated sequence.

Month (m)	16	17	18	19	20	21	22	23	24
Concepts acquired in month m (n_m)	17	3	15	9	4	6	14	9	6
Cumulative concepts (N_m)	17	21	36	45	49	55	68	77	83

Table 2.1: Number of concepts acquired in each month, based on mean number of concepts known in each month of the WordBank dataset. Values are rounded to the nearest whole number of concepts.

Forced choice task

To test the extent to which agents' knowledge states facilitated learning new concepts by alignment, we used a forced-choice paradigm. In this paradigm, simulated agents attempted to infer the correct mappings for a novel pair of probe concepts presented in word and image embedding spaces. Probe concepts were sampled using two different *probe conditions*: either from the remaining AoA concepts (AoA-constrained probe condition) or from all remaining concepts (Unconstrained probe condition).

Agents made their choice by assessing the *alignment score* for the two possible image-word mappings—one of which mapped each probe word to its correct visual object, and the other of which yielded the incorrect mapping (visualised in Figure 2.7).

The scoring procedure for the forced choice task is visualised in detail in Figure 2.8. The alignment score measures the extent to which within-

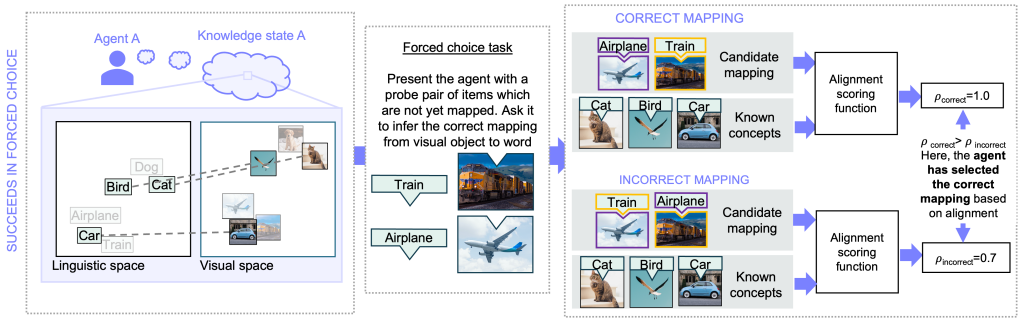


Figure 2.7: Example of the forced choice task used to evaluate agents. In this example, Agent A's knowledge state allows it to make the correct inference in the forced choice task, using alignment. In the left panel, the agent's knowledge state prior to the forced choice task is represented. Greyed out images and words in the visual/linguistic spaces represent items which the agent has experienced separately in each modality, but which have not been mapped across systems. The next panel to the right shows an example forced choice task: in this case, agents are asked to infer which of two visual objects is an 'Airplane' and which is a 'Train'. The next panel shows how the agent attempts this inference. The agent obtains the alignment score for each candidate mapping of the probe items using the alignment scoring function. The alignment scoring function is shown in detail in Figure 2.8. Agent A correctly identifies the appropriate mapping, because the alignment score for the correct mapping is higher than the score for the incorrect mapping. For an example of a knowledge state which would yield failure in this forced choice task, see Appendix A1.

system similarity relationships correlate for corresponding items in a given cross-system mapping. Given that the mapping for all but the probe items was fixed in the forced choice task, the higher alignment score can be determined by the Spearman correlation, ρ_s, between concatenated pairwise distance vectors for the probe items' distances from the known concepts in each modality. The order of concatenation in each modality was determined by the proposed cross-modal mapping (see Figure 2.8). A forced choice was deemed correct if the correlation for the correct mapping was higher than for the incorrect one.

Both probe conditions are tested on both agent conditions, with 100 simulated agents for each condition. Month m and probe condition are within-subjects factors; agent condition is a between-subjects factor. This yields a two-way repeated measures design.

2.3.2 Results

The results are shown in blue (AoA) and red (Control) in Figure 2.9. First, it is striking that with only a handful of known concepts that both agents' inferences are over 80% accurate in the forced choice task (see Table A.3 for month-wise t-tests). This indicates that children, like our agents, could align systems (e.g., visual and words) to correctly label objects using similarity re-

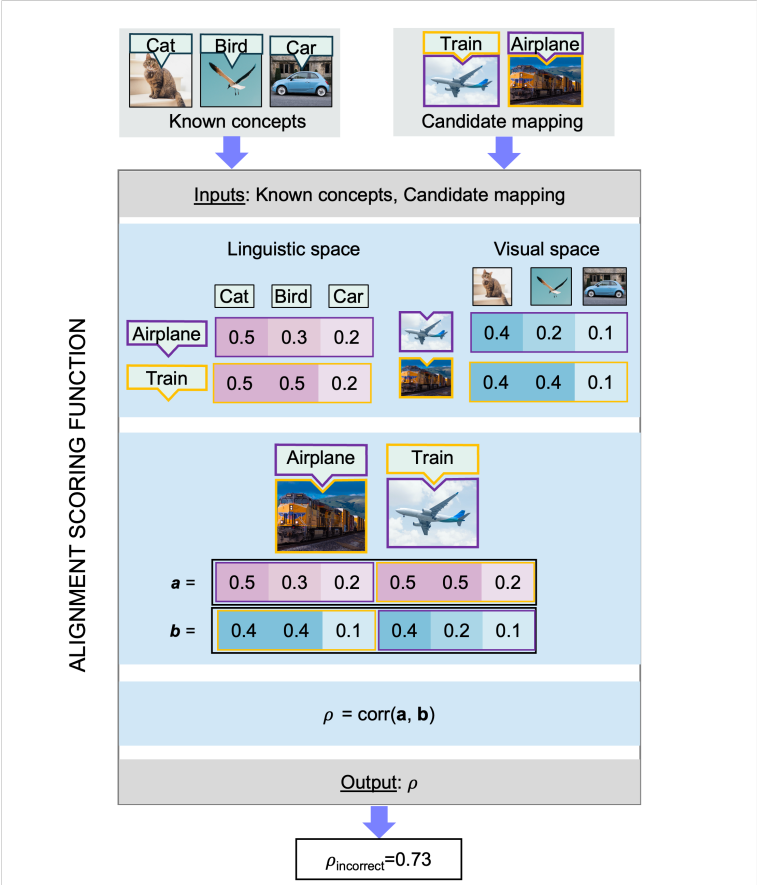


Figure 2.8: Details of how the score is calculated for a candidate forced choice mapping, using an agent's knowledge state. First, the agent retrieves the inter-concept distances for the probe items with respect to its known items in each modality. Then, the similarity relationships in each modality are concatenated in the order determined by the candidate mapping. The resultant vectors are correlated across modalities. The chosen mapping is the one which maximises the correlation between similarity vectors across modalities

lationships to their known concepts. These results extend those of Roads and Love (2020) to suggest that systems alignment is useful for inferring unknown concepts and building useful priors or expectations.

Second, we found that the AoA agent was more effective than the control agent ($F(1, 198) = 347.48, p < .001, \eta_p^2 = 0.627$) and an advantage of AoA test probes ($F(1, 198) = 529.96, p < .001, \eta_p^2 = 0.719$). There was also a significant interaction between agent and probe type ($F(1, 198) = 11.83, p < 0.001, \eta_p^2 = 0.069$) such that the AoA probe effect was heightened for the AoA agent. Complete ANOVA results are provided in Table 2.2.

2.4 Analysis of structural features

Based on the findings in the forced-choice experiment, we hypothesised that early-acquired concept sets possess common structural features which facili-

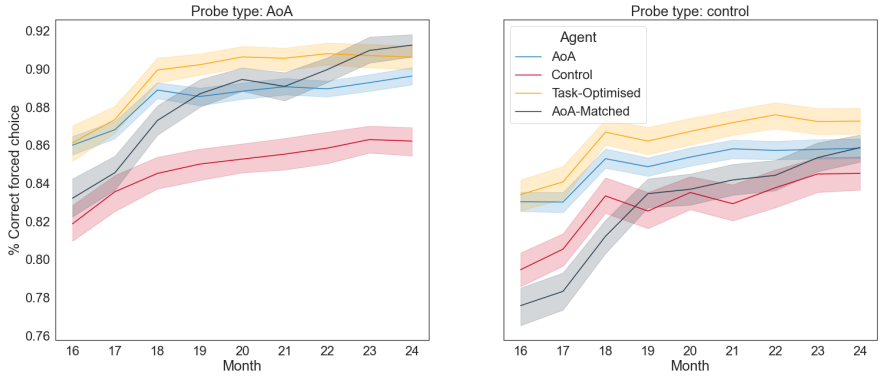


Figure 2.9: Results for forced choice experiment for different agent types. Shaded regions represent 95% confidence intervals across 100 agents for each agent type. **AoA vs control** Blue lines represent performance for agents simulating AoA-based concept acquisition, and red lines represent results for control agents. **Generative modelling** Orange and black lines represent results for structural agents. Black lines represent performance for the agents which are trained to match AoA acquisition statistics; orange lines represent performance for the agents which are trained to optimise probe pair performance.

Predictor	df	F	p	η_p^2
Agent	$(1, 198)$	347.48	$< .001^*$	0.627
Probe	$(1, 198)$	529.96	$< .001^*$	0.719
Agent * Probe	$(1, 198)$	11.83	$< .001*$	0.069
Month	$(8, 1584)$	62.11	$.001^*$	0.235
Agent * Month	$(8, 1584)$	2.61	$.008^*$	0.011
Probe * Month	$(8, 1584)$	1.84	.066	0.008
Probe * Agent * Month	$(8, 1584)$	1.09	.369	0.005

Table 2.2: Repeated-measures ANOVA results for probe pair experiment. Agent condition (AoA vs control) and probe condition (AoA-constrained vs Unconstrained) were between-subject factors and month was a within-subject factor. $*$ indicates statistically significant results for α=0.05. df = degrees of freedom; η_p^2 is partial η^2 effect size.

tated the observed uplift in learning by alignment. In this analysis, we sought to identify quantifiable structural features which distinguish an AoA knowledge state from a Control knowledge state, and aimed to explore whether these features drove alignment uplift.

We calculated a range of features for the knowledge states of both Control and AoA agent types. The different features were derived from similarity relations of concepts and graphs of concepts' close neighbours. Where applicable, these features were calculated for each concept with respect to both (a) the full space of all concepts, and (b) the set of concepts already in the agent's knowledge state at the point of acquisition. Additional features characterised the knowledge state as a whole, including its average coverage of embedding space dimensions and the distribution of node degrees within the knowledge state (see Table A.4 for the full table of features tested).

Most tested features are averages of concept-wise features taken across the concepts in the knowledge state. These fall into one of two broad categories:

- **Global similarity features:** These are features based on similarity relationships in the full system of concepts. These features are rooted in the similarities between each concept and others in the system. For example, the mean global distance for a concept i would be the mean of i's distance to every other concept in the system.

- **Neighbourhood graph features:** These are derived from graphs constructed from only the shortest-range inter-concept relationships (or, in other words, concepts' immediate neighbourhoods). We build a graph G, whose nodes are concepts within an embedding space, by retaining the vertices for the 10% of smallest inter-concept distances, based on the similarity matrix.

Note that the graphs we generate in this study are not necessarily connected, therefore some graph measures such as smallworldness are not applicable in our case. Clustering and betweenness measures were obtained using **networkx** in Python (Hagberg et al., 2008). While we explored clustering and betweenness results for the AoA vs control knowledge states, we excluded these variables from selection by the logistic regression model due to the computational demands of calculating them for model training. This exclusion had no impact on the performance of the model selected. All features were normalised to fall in [0, 1].

A logistic regression classifier was trained to predict if a knowledge state was sampled under the AoA or the Control condition. Logistic regression was performed using **scikit-learn** in Python (Pedregosa et al., 2011). An 80/20 training/test split was applied to the knowledge states in advance of training. Logistic regression was performed with a *liblinear* solver and L2 loss, where the maximum number of iterations was set to 10,000.

We used recursive feature elimination to identify the features which were most powerful in demarcating early-acquired knowledge sets. When applying

recursive feature selection, values of k (number of features included in the model) from 1 to the full feature set were tested, and the value which minimised each model's *Akaike Information criterion (AIC)* was selected.

Regression results are shown in Table 2.3. Recursive feature selection chose seven features: distance to closest neighbour in full space and knowledge state, mean degree in full space and knowledge state, degree distribution skew in knowledge state, mean distance in knowledge state and mean dimension coverage of the knowledge state. The accuracy, recall and precision the resultant model were all over 92% for a balanced set of 900 Control and 900 AoA knowledge state samples, meaning that all models correctly classified a significant majority of samples on the basis of these features.

Feature	β
$\text{Degree}_{\text{knowledge}}$	7.21
$\text{Degree}_{\text{full}}$	5.36
$\text{Mean}(\text{Dist}_{\text{knowledge}})$	-3.28
Mean dimension coverage	-7.43
$\text{Min}(\text{Dist}_{\text{knowledge}})$	-4.04
$\text{Min}(\text{Dist}_{\text{full}})$	-2.60
$\text{Skew}(\text{Degree}_{\text{knowledge}})$	4.76

Table 2.3: The β values of logistic regression after recursive feature elimination. The regression model was trained to classify sample knowledge states as early-acquired ($Y = 1$) or control ($Y = 0$).

The regression analyses indicated that AoA concepts are distinguished by their dense neighborhoods. From a systems alignment perspective, density may be advantageous because it promotes stability. Embedding algorithms are sensitive to initial conditions such that the position of items within an embedding can vary across simulations. Human learners may also be affected by these and other factors, such as noise and the idiosyncratic nature of human experience. We confirmed the stability hypothesis: longer-range relationships are less stable across embedding initialisations (see Figure 2.10), meaning that dense neighborhoods characteristic of AoA concepts are better suited to system alignment. Stability may explain why children preferentially acquire concepts with many semantic neighbours (Hills et al., 2009; Stella et al., 2017).

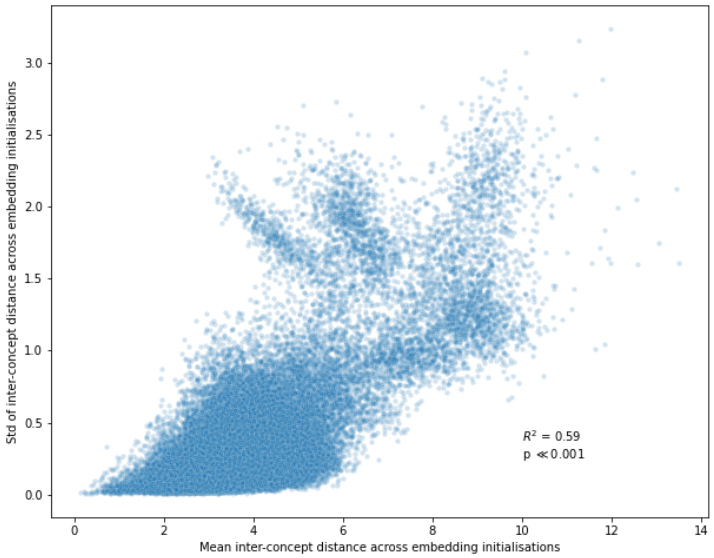


Figure 2.10: Relationship between the average distance between two concepts and the standard deviation of the relationship across multiple initialisations of the embedding space.

2.5 Learning with generative agents

Having identified structural features of early-acquired concepts within embedding spaces, we explored whether these features could be used to build knowledge states which are optimal for alignment. Could removing the additional constraints which apply to early concept acquisition yield even better alignment performance, using the same structural features that AoA concepts seem to favour? We introduced two new agents, which we refer to as *structural agents*. We began by testing the forced-choice performance of a structural agent which was trained to predict empirically-derived knowledge states (henceforth the *AoA-Matched* agent). We also trained a separate *Task-Optimised* agent, which aimed to maximise performance on the forced-choice task. The Task-Optimised agent serves as an upper bound on the forced-choice performance which can be achieved by optimising these structural parameters, to which the AoA-Matched agent's performance can be compared.

2.5.1 Methods

The training and generative processes for structural agents are outlined in Figure 2.11, and are described in detail below.

Model structure

Both agents were set up to learn a vector of target values $\hat{\mathbf{x}} \in \mathbb{R}^k$ for the structural features we identified as being predictive of early acquisition, where k is the number of features available to learn ($k = 7$). Agents also learned a weight vector $\mathbf{w} \in \mathbb{R}^k$, which captured the relative importance of each feature.

Generating knowledge trajectories based on model parameters

For each month m in the AoA data, the agent samples new concepts one at a time from probability distributions generated based on the current model parameters. These probability distributions are derived from the proximity of the candidate knowledge state associated with each candidate concept to the current target values $\hat{\mathbf{x}}$ in structural feature space, weighted according to the current estimate of feature importances. This is shown in panel 2a of Figure 2.11.

When selecting a new concept to add to the knowledge state according to our current model, we construct a generative score vector for the candidate concepts $\mathbf{s} \in \mathbb{R}^n$ (where n is the number of candidate concepts for acquisition) as follows:

- Obtain feature matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$ for all n concepts in candidate concept set (i.e., all concepts which have not yet been acquired.)
- Get generative scores for all concepts, $\mathbf{s} = -\mathbf{w}^T(\mathbf{A}^T - \hat{\mathbf{x}}\mathbf{J}_{1,n})$, where $\hat{\mathbf{x}} \in \mathbb{R}^p$ is the model's current best estimate of target feature values, $\mathbf{w}$ is the current estimate of feature weights and $\mathbf{J}_{1,n}$ is an $1 \times n$ matrix of ones. The generative score is highest for sampled items whose feature values have the smallest distance to the current target feature values, weighted by feature importance.

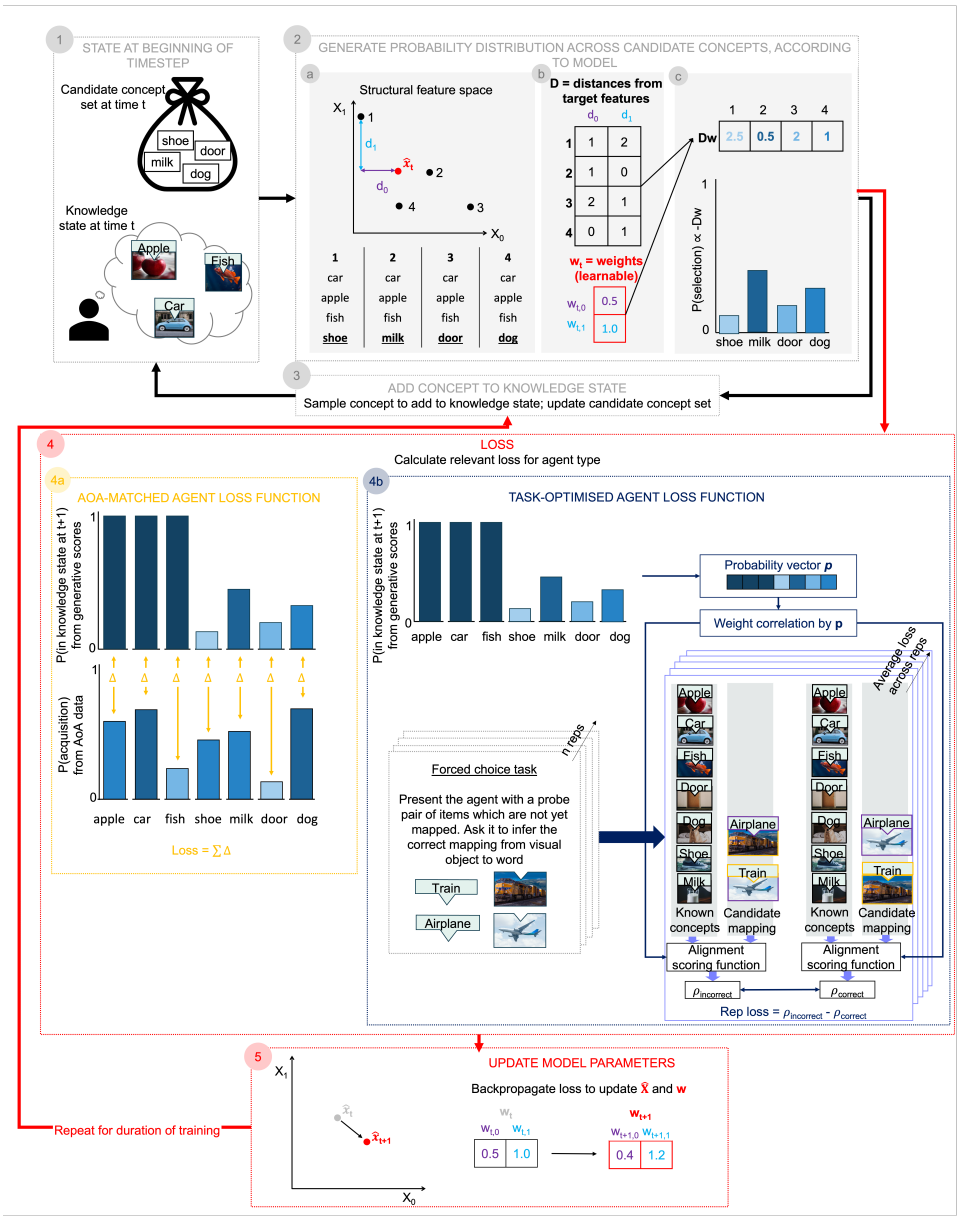


Figure 2.11: Diagram showing the training and generative processes of the structural agents. Model parameter training proceeds as follows: assume we start observing training when the knowledge state and candidate concept set are as shown in panel 1. Here, the agent knows {apple, car, fish}. Panel 2 shows how the agent uses its internal model to calculate a probability distribution for the selection of the next concept from the n_c candidates shown in panel 1. The internal model consists of target structural feature vector $\hat{\mathbf{x}}_t \in \mathbb{R}^k$ and weight vector $\mathbf{w}_t \in \mathbb{R}^k$. For visualisation purposes we show the case of $k = 2$, but in the main study $k = 7$ (as 7 structural features were identified in the structural analysis). Internal model parameters and model training steps are highlighted in red. In panel 2a, all structural features are calculated for each candidate knowledge state (where, for example, the candidate knowledge state associated with acquiring the concept 'shoe' is {apple, car, fish, shoe}). Then, the distance of each candidate knowledge state from $\hat{\mathbf{x}}_t$ in each dimension is calculated, yielding distance matrix $\mathbf{D}^{n_c \times k}$, shown in panel 2b. The vector of weighted distances from target in each dimension, $\mathbf{Dw}$, is calculated in panel 2c for all candidate knowledge states. This is transformed into a probability distribution across candidate concepts, where candidate knowledge states with features close to $\hat{\mathbf{x}}_t$ are chosen with higher probability. In step 3, the next concept for the knowledge state is sampled from this distribution. In model training, the agent progresses to step 4 at the end of each month m, where it calculates the relevant loss for optimisation. For the AoA-matched agent (shown in orange, 4a), the loss is the distance between the expected probability that each concept will be in the knowledge state according to the model, and the probability of each concept being acquired by the end of month m in the AoA data. In this instance, the acquisition of each concept is modelled as an independent Bernoulli random variable. For the Task-optimised agent (shown in blue, 4b), the loss is $\rho_{\text{incorrect}} - \rho_{\text{correct}}$, averaged across a series of forced choice tasks. Crucially here, the correlations are weighted by the probability that each concept will be selected for the knowledge state. Then, in step 5, the loss term is backpropagated to update parameters $\hat{\mathbf{x}}$ and $\mathbf{w}$. Once a model is trained, parameters are fixed and knowledge state trajectories are generated by repeating steps 1, 2 and 3.

- Generate probability distribution across candidate concepts by taking the softmax of normalised scores (softmax temperature parameter $T = 5 \times 10^{-2}$). The closer a candidate knowledge state is to the target values in the weighted feature space, the more likely the associated candidate concept is to be selected.

- Sample a concept from this probability distribution and add to knowledge state.

Loss and model training

At the end of each month in the data, where the number of acquired concepts in simulation matches the average number of concepts acquired from the WordBank dataset, we backpropagate our loss. The key distinction between the AoA-Matched and Task-Optimised agents was the loss function that is backpropagated to optimise the agent's internal parameters. The AoA-Matched loss term pressured the agent's probability of acquiring each concept to match the real probability of acquisition from the AoA data. Meanwhile the Task-Optimised agent aimed to directly optimise performance on the forced-choice task by using a loss term which pressurised the model to maximise the margin between alignment scores for correct and incorrect mappings ($\rho_{\text{s, correct}} - \rho_{\text{s, incorrect}}$) across a randomly selected set of forced-choice problems. Details on both loss terms are provided below.

AoA-Matched agent The loss for this model is the average MSE between (a) the model's estimated probabilities of each concept being included in the knowledge state by the end of month m and (b) a set of bootstrapped probability distributions sampled from the WordBank acquisition probabilities. The model's estimate of the probability that an item it has already selected for its knowledge state being in the knowledge state by the end of month m is set to 1. The probabilities for remaining candidate concepts are determined by the probability distribution across candidates outputted by the current model. We

train $R = 5$ model restarts on average MSE across training set bootstrapped distributions. Models are trained for 150 epochs. An Adam optimiser with a learning rate of 0.003 was used for training. We select the best model based on validation loss averaged across the final 5 epochs.

To obtain ground truth probability distributions for AoA-Matched training, we take bootstrap samples from the probability distributions of concept acquisition in WordBank. Details are provided in Appendix A5.

Task-Optimised agent To train the optimal model, we backpropagate a *soft alignment loss* across a sample of probe pairs, where the alignment loss is the extent to which the incorrect alignment score is greater than the correct alignment score, averaged across pairs. The larger the margin for the correct alignment score (i.e, the clearer the correct answer is), the smaller the loss becomes. The alignment loss is *soft* because it is weighted by the candidate concepts' probabilities of being selected for the knowledge state. The process of calculating the soft alignment loss is presented in Appendix A6.

As before, we train $R = 5$ model restarts to minimise this alignment loss, and models are trained for 150 epochs. An Adam optimiser with a learning rate of 0.01 was used for training. We select the best model based on validation loss averaged across the final 5 epochs, where validation loss is soft alignment loss calculated on a validation set of forced-choice items.

Testing resultant knowledge states

After each agent model was trained, model parameters were fixed and agents were tested. Using the learned model parameters, agents simulated concept acquisition across an eight-month period. Trajectory generation occurred in the same way as it did during model training, but without parameter updates. Each model was tested using 100 generative simulations in order to establish a reproducible result. For each simulation, agents sequentially expanded their knowledge states by selecting concepts that satisfied the agent's previously learned feature weightings (e.g., an outlier concept was selected because it

better satisfied the agent's preference for wider coverage of the embedding space). As the agent sequentially selected concepts, it was tested each month using forced-choice tasks, according to the same procedure used for AoA and Control agents in Experiment 1.

2.5.2 Results

Forced choice results

Forced-choice results for the generative paradigm are visualised in Figure 2.9. Results for the AoA-Matched and Task-Optimised agent are shown in black and orange respectively. Significance tests for all agent comparisons are provided in Table A.5.

When the forced-choice task involves AoA concepts, the AoA-Matched agent performs worse than AoA Agent in the first three months and outperforms the AoA agent in the final months. When the forced-choice task involves Unconstrained concepts, the AoA-Matched agent performs worse than the AoA agent during early months, but performs at a similar level in later months. The Task-Optimised agent performs better than the AoA-Matched agent in the early months (for both probe types) and performs better than the AoA agent for later months (for both probe types).

Structural trajectory analysis

An analysis of the proportions of early-acquired concepts in the knowledge states learned by each agent type demonstrates that the performance of structural agents did not rely solely on concepts in the AoA set, with 57% and 62% of concepts being early-acquired in the final knowledge states of Task-Optimised and AoA-Matched agents respectively (see 2.12). Naively, one might think AoA trajectories are the only learning path which yields success using these structural features. However, these results demonstrate that there is a more general solution space which lies outside of what is observed empirically. In the final knowledge state, where the Control condition consisted

of only 33% AoA concepts, both AoA-Matched ($t(198) = 48.9, p \ll 0.01$) and Task-Optimised agents ($t(198) = 43.7, p \ll 0.01$) demonstrated enhanced preference for AoA concepts compared to Control agents. This was true despite the Task-Optimised agent having no explicit training pressure to select AoA items.

It is interesting to note that the proportion of acquired items which were in the early acquired set in each month declined over time (see Figure 2.13): the structural agents started out by choosing concepts which were somewhat similar to those chosen by an AoA agent (indeed, nearly 70% of the concepts each agent chooses are early-acquired in the first month), but both agent types increasingly branched out into non-AoA concepts over time. AoA-Matched agents did not consistently acquire a higher proportion of early-acquired concepts than the Task-Optimised agents. This goes to show that while there are clearly additional constraints on early concept acquisition which were not captured by our structural features—resulting in the number of non-AoA concepts acquired by AoA-Matched agents—early-acquired concepts do indeed possess a privileged position when it comes to optimising for alignment.

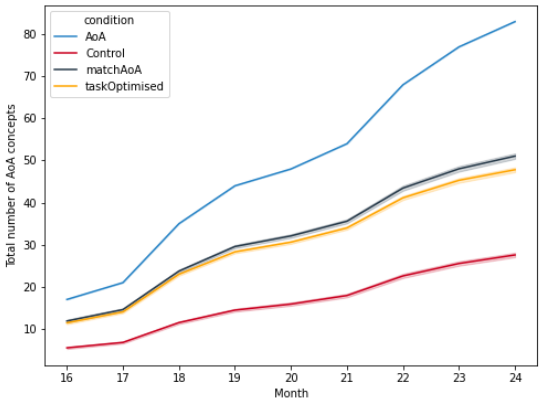


Figure 2.12: Number of concepts acquired which are early-acquired, by generative condition. The AoA line represents all items in the sample being early-acquired.

The two agent types had different priorities when selecting new concepts for acquisition: the Task-Optimised agent, which performed better in forced-choice, prioritised learning concepts which had many close neighbours in the full semantic space; the AoA-Matched agent, on the other hand, prioritised

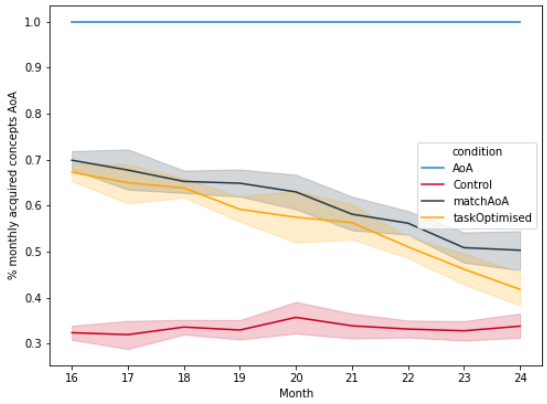


Figure 2.13: Proportion of concepts acquired in each month which are in the set of early-acquired concepts found in WordBank.

acquiring concepts which had low mean distances from other concepts in the existing knowledge state and many close neighbours within the knowledge state (see Figure 2.14). This suggests that Task-Optimised agents achieved their superior performance by focusing on concepts which have dense similarity neighbourhoods in semantic space. AoA-Matched agents were prone to select knowledge states with low coverage, as was seen in the classification results, but this was not true of Task-Optimised agents, indicating that while low coverage may be a feature of early-acquired knowledge states, it does not necessarily contribute to the enhanced alignment effect.

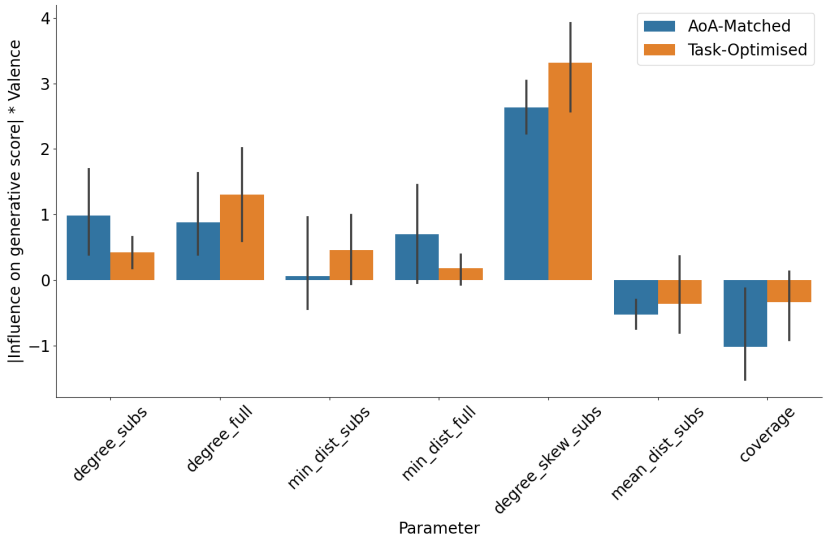


Figure 2.14: Mean learned importances of features for selecting new concepts to add to the knowledge state, for each generative agent type. Error bars represent 95% confidence intervals across restarts.

So, what types of concepts do these structural features lead agents to seek

out? We examined the semantic category coverage of the concepts learned by each agent type (see Figure 2.15). The AoA-Matched agents show a similar distribution across categories to the AoA agents, while the Task-Optimised agents have a tendency to focus in on fewer semantic categories, as indicated by low cross-category entropy. This implies that the Task-Optimised agents prefer specialism or depth of knowledge within categories, as opposed to covering all category bases as a priority for knowledge acquisition.

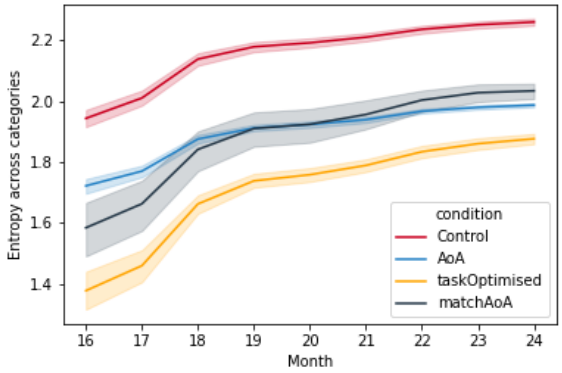


Figure 2.15: Overall entropy of knowledge state's category distribution after each month of concept acquisition. Shaded areas represent 95% confidence intervals around the mean, based on 100 simulated agents per condition.

2.6 Discussion

This chapter has demonstrated that aligning systems can aid the acquisition of conceptual knowledge. By matching inter-concept relationships alone, it is possible to infer the meanings of concepts with over 80% accuracy in a forced choice task with a knowledge state containing only 21 known concepts. This startling result, along with prior work on systems alignment (Roads and Love, 2020) and supportive brain imaging findings (Popham et al., 2021), suggests a revised account of human learning. Rather than being strictly event-based, human learning may also draw upon the alignment of conceptual systems, which can link asynchronous events in an offline manner, akin to how neuroscientists characterise memory replay (Barry and Love, 2023).

We found that children's early-acquired concepts provided knowledge states that better supported alignment than randomly-sampled knowledge states

(i.e., AoA vs Control agent results, Figure 2.9). In turn, early-acquired concepts were easier to learn by alignment than later-acquired concepts (i.e., AoA vs Unconstrained probe type results, Figure 2.9). These findings suggest that children could engage in the types of unsupervised learning involved in systems alignment, which would lead to a preference for concepts forming alignable systems. A complementary possibility is that children are biased to acquire alignable systems of concepts based on some structural property of these knowledge states.

In accord with this second possibility, we found that children's early (AoA) concepts were distinguished from other concepts by certain structural features, such as being densely packed and interconnected (see Table 2.1). We predicted that these features were particularly beneficial for systems alignment. To evaluate this possibility, we built agents that used these features to select concepts to learn through systems alignment (see Figure 2.7). As predicted, these agents were more effective than agents that randomly sampled concepts (see Figure 2.9). The AoA agent patterned after children's acquired concepts performed nearly as well as the Task-Optimised agent which indicates that children's early concepts provide a knowledge base highly suited to (and perhaps shaped by) systems alignment.

One factor underlying the success of these agents (and children) may be that dense semantic neighbourhoods provide a solid foundation for subsequent concept learning. In support of this conjecture, we found that relationships in embeddings across multiple initialisations are most stable for shorter-range relationships, which would make knowledge bases consisting of densely packed concepts most reliable for systems alignment. This inherent sensitivity to initial conditions and noise in learning systems may privilege densely packed concepts as found in children's early concepts.

These structural agents were constrained to follow the feature patterns characteristic of children's early concepts rather than learn the specific concepts children do. Their success demonstrates that there are multiple paths

to successful learning by alignment. Indeed, the specific concepts learned by the structural agents differed from those children learned (see 2.12). When we use our structure-based models to generate sequences using only concepts which are not in the AoA dataset at all, they still achieve forced-choice performance with up to 90% accuracy at the maximum knowledge state size tested (see Figure A.4). In summary, while children's early concepts form a readily aligned system, there are many other knowledge states that also support systems alignment.

Unlike the artificial agents, children likely face a trade-off between the concepts which are easiest to integrate with their current knowledge by alignment -i.e., those with dense connections- and those which they must learn in order to gain a functional understanding of their world. We found evidence of this trade-off by comparing the structural preferences of agents which were trained to mimic early-concept acquisition and agents which were trained solely to optimise alignment performance. Agents which mimicked real-world concept acquisition struck a balance between densely-connected knowledge states and knowledge states which spanned semantic categories, while agents optimising alignment performance honed in on a narrow range of semantic categories, favouring connection density.

We focused exclusively on systems alignment, which we view as an exciting and under-explored avenue for learning. We fully acknowledged that other factors shape early concept acquisition. Supervised episodes, and other aforementioned event-based inputs, undeniably affect children's learning. Rather than be in opposition, systems alignment is compatible with other forms of learning, including event-based learning. Children's learning likely reflects a mix of systems alignment and event-based learning. In our own results, we found that AoA concepts tend to be higher frequency, which we take as a marker of event-based learning. Thus, children's early concepts show indications of both systems alignment and event-based learning. When we limit our simulations to non-AoA concepts that tend to be lower frequency, systems

alignment continues to perform well (see figure A.4), suggesting that it may be possible to disentangle these forms of learning that are likely intertwined in natural environments.

Systems alignment can explain how learning is possible from weak supervisory signals. In naturalistic environments, weak signals may come in the form of ambiguous (Quine, 1960; Yu and Smith, 2007) or infrequent (Clerkin and Smith, 2022; Karmazyn-Raz and Smith, 2022) labelling events, or indeed from the context-specificity of early language (Tamis-LeMonda et al., 2019; Roy et al., 2015). These signals could constrain systems alignment processes by suggesting links between systems and restricting the set of candidate solutions. In turn, systems alignment could help constrain weakly-supervised mapping problems by favouring mappings that mirror similarity relationships across systems.

Learning via systems alignment remains to be tested in children under controlled conditions. Our results invite directed laboratory studies to evaluate whether children's learning is accelerated by systems alignment. Like our agents, we predict children should be able to infer novel mappings between objects and labels using systems alignment.

We made the simplifying assumption that each embedding or similarity space is constant over time. While there is evidence that similarity spaces apply over development – co-occurrence statistics derived from child-directed speech generate adult-like word embeddings (Li et al., 2000; Unger et al., 2020b) – and our own analysis showed that the alignment benefit for early-acquired concepts is also observed when using child-directed speech embeddings (see SI text), one would expect some changes in these spaces over learning. Infant environments, while certainly correlated with adult environments for the concepts they are exposed to, are more constrained than adult environments, and semantic spaces will develop over time.

A limitation of this work remains in the fact that the visual object embeddings are not specific to children's environments. Although the similar-

ity relationships between the concepts studied here are likely to be largely re-capitulated in such embeddings, these embeddings may align even better with child-directed speech embeddings. Future work could infer more child-like visual representations to explore how alignment signals manifest in early similarity space.

Additionally, while our visual embeddings are based on visual object co-occurrences, the semantics of the visual space may instead be captured by embedding other kinds of visual information, such as how objects co-occur with actions and contexts (Tamis-LeMonda et al., 2019; Roy et al., 2015) or objects' perceptual features (Riordan and Jones, 2011). Embeddings based on these types of visual information may better capture how children judge visual similarity, which may improve systems alignment to linguistic spaces (Johns and Jones, 2012).

To simplify, we considered conceptual understanding as 'all-or-nothing': when a word-image mapping is known, the concept is 'understood'. Instead, conceptual understanding, and indeed cross-modal mappings themselves, are likely graded. One possibility is that systems alignment may provide informative priors for concept learning, thus facilitating event-based learning. For example, possible alignments that lead to higher alignment scores could be assigned higher priors. In turn, event-based learning constrains systems alignment by expanding the knowledge base. These principles could benefit machine learning systems using alignment-informed priors for multi-modal learning.

This chapter has demonstrated that the concepts children learn in early life are particularly well-positioned to support learning by systems alignment. A natural question to follow is whether people do in fact benefit from alignable systems when learning. As noted here and in the previous chapter, there are a host of established learning signals which are known to contribute to concept learning. So, more specifically, do alignment signals bolster human learning in the presence of other learning signals? The next chapter of this thesis addresses this question via a behavioural study. The answer to this question will help

to understand the hypothesised role of alignment in resolving ambiguity in real-world learning signals.

Chapter 3

Alignment in supervised learning

Prior work (Roads and Love, 2020; Johns and Jones, 2012) and the simulation study presented in the previous chapter converge on the theoretical utility of alignment-based signals for learning cross-modal mappings. But while these signals have been shown to exist in naturalistic learning environments, do humans capitalise on alignable systems in learning? In this chapter I explore this phenomenon in a human-subject experiment, which tests whether cross-system learning is supported by shared structural relationships. We investigated whether participants were better able to learn associations between *aligned* systems than *misaligned* ones in the presence of supervised learning signals.

Learning is often viewed as event-based. For example, pairing a face with a label provides a means to learn a stranger's name. A complementary possibility is that humans learn by establishing correspondences between entire *systems* (Goldstone and Rogosky, 2002).

Imagine you are abroad with your partner who is watching a basketball game on television in an unknown language. You are facing away from the television unpacking your luggage. You frequently hear cheering followed by the announcer saying various utterances containing "Michael". Your partner, noticing your disinterest in the game, plugs their headphones into the televi-

sion. Turning toward the muted television, you notice the same star player from the home team keeps scoring. Despite being limited to asynchronous cross-modal input, a reasonable inference based on aligning systems is that the star player's name is Michael.

Mappings like this are possible far beyond simple features like frequency. For instance, similarity relations across visual and linguistic systems may mirror one another: cups and mugs appear in related linguistic contexts concerning drinking and also are visually similar.

We have presented evidence that the information exists in the real world to support aligning conceptual systems based on similarity relations. To reiterate, Roads and Love (2020) conducted an information analysis on different unimodal embeddings, which found that similarity relations remain consistent across modalities. That is, if 'car' and 'truck' occur in similar linguistic contexts, their corresponding referents are likely to occur in similar visual contexts (Figure 1.1). The previous chapter demonstrated that these relationships are particularly prevalent in early-acquired concepts, and could support successful concept learning in early life.

We define a *system* as a set of items organised within a *domain*, where a domain is the set of possible inputs to a mapping function $F(X)$ for a given task (see Figure 1.1). In learning to label visual objects, we learn correspondences between systems of items within visual and linguistic modalities. These modalities are the domains, in this case[1]. While perceptual modalities are clearly relevant examples of domains between which humans regularly establish correspondences, this chapter aims to demonstrate a domain-general process through which humans may capitalise on cross-system structure to boost learning.

Research in analogy seeks alignments between representations (Gentner, 1983; Holyoak and Thagard, 1989; Lu et al., 2012; Doumas et al., 2019), but whereas analogical alignment is between two analogs, such as an atom and

[1]Domains could also be contained within a single modality: in translating between two languages or performing analogy, correspondences may be established between different systems of items within the linguistic modality.

the solar system, we suggest that entire conceptual systems could be aligned. Systems alignment also diverges from alignment work in category learning (Lassaline and Murphy, 1998) and in similarity perception (Goldstone and Medin, 1994), as it does not require features to be shared across systems for mapping, and depends instead on the similarity relationships within systems.

As described in the previous chapter, systems alignment offers a possible explanation for humans' remarkable success in acquiring multimodal concepts, despite this being a famously challenging and underconstrained task. Infants can acquire an understanding of more than 300 concepts by 16 months of age (Fenson et al., 1994). Yet even the most *supervised* learning episodes—such as pointing at an object while naming it aloud—are ambiguous. This problem of referential ambiguity is demonstrated by Quine's famous thought experiment (Quine, 1960); if a teacher points at a rabbit hopping through a field and says 'gavagai' aloud to a naive learner, how does the learner know what 'gavagai' refers to? It could mean hopping, rabbit, fur, field - the list of possibilities goes on.

To solve this problem, systems alignment could facilitate cross-modal learning *offline* (that is, in the absence of synchronous input across systems) by capitalising on common structural relationships. For example, the systems in Figure 1.1 could be mapped by matching the similarity relationships between concepts across domains, requiring no synchronous input across modalities. As such, systems alignment can explain learning from ambiguously supervised events (such as those discussed in the 'gavagai' problem), and even in the absence of explicit instruction (Cartmill et al., 2013; Lieven, 1994; Samuelson et al., 2011).

While systems alignment enables purely unsupervised learning, signals about the strength of alignment may also enhance learning in the presence of supervised examples, as memory of individual item mappings is reinforced by the alignment of systems. In this study, we aimed to investigate whether participants were better able to learn associations between *aligned* systems com-

pared to *misaligned* ones in a supervised learning task (Figure 3.1). Aligned systems are those for which the correct pairing of objects between systems is dictated by their second-order isomorphism. This means paired items share a pattern of relationships within their respective systems, while sharing no physical properties (Shepard and Chipman, 1970). In a misaligned set of systems, paired items share neither physical properties nor patterns of relationships.

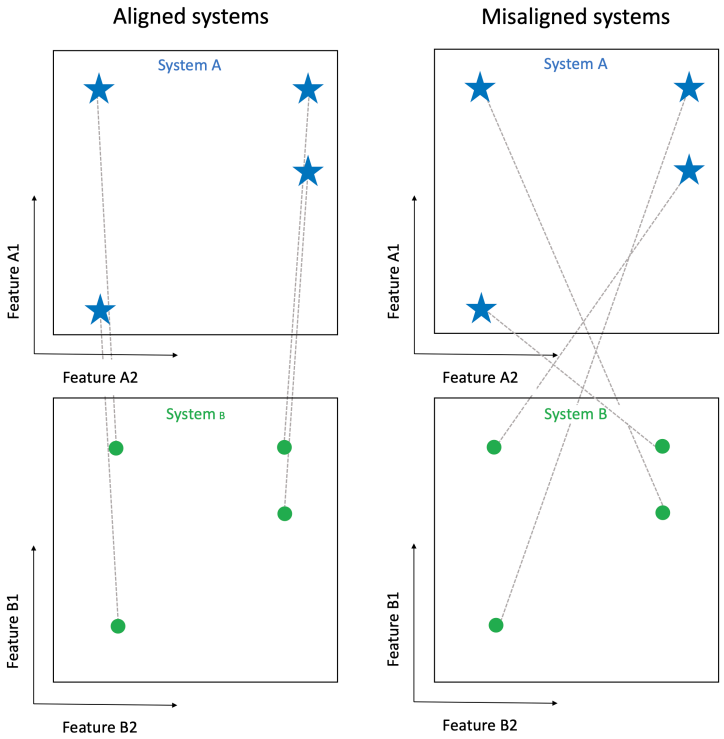


Figure 3.1: Examples of aligned and misaligned systems. In aligned systems, similarity relations are recapitulated across systems, which is not true for misaligned systems.

Our primary hypothesis is that learning will be facilitated when systems align, even in cases where feedback is provided and synchronous. That is, even when systems alignment is optional for success in the learning task, people will engage in it. A default systems alignment strategy might produce idiosyncratic error patterns for misaligned scenarios.

Systems alignment should create expectations for how an unseen example maps from domain X to Y based on its relationships to other items in X (Figure 1.1). This can be described as *zero-shot generalisation* (Xian et al., 2017): unlike classic generalisation tests (e.g in category learning), zero-shot generalisation could occur where items in domain X and domain Y are both

novel, provided that their relationships to other items in the system were known. We predict that participants who align systems should be able to perform zero-shot generalisation to a novel stimulus in X to Y, which would be like knowing the name of visual object one has never encountered before. Finally, we predict that a computational model including an offline alignment mechanism would be the best fit for participants in the aligned condition, compared to models simulating (i) rote-memorisation and (ii) cross-system mapping with no distributional alignment.

3.1 Experimental Methods

3.1.1 Design

We tested the primary hypothesis using a paired-associate learning (PAL) paradigm, presented as a memory game. Participants were tasked with learning where a set of cartoon monsters lived on a map across a series of trials. This is a similar procedure to the 'image-location association' procedure used in Tompary and Thompson-Schill (2021).

The monsters varied on two feature dimensions: body colour and eye orientation, where their eye was an orientation grating. In the aligned condition, the relationships between monsters based on these two features could be mapped onto the relationships between their corresponding houses as shown in Figure 3.2.

To account for the possibility that alignment effects could result solely from the mapping of feature dimensions onto the privileged set of canonical axes, the experiment also included a rotation condition. This was used to explore whether the impact of alignable systems varied based on the rotation of the spatial axes.

The experiment proceeded with a 2x2 (system alignment x rotation) repeated-measures design across 5 blocks of 6 trials each, for a total of 30 paired-association trials per participant. One further trial tested generalisation to an

unseen stimulus. Participant assignments across the four experimental conditions were counterbalanced.

3.1.2 Neighbourhood stimuli

The rotation condition was included to control for the possibility that participants could align privileged axes instead of whole spaces. The relative positions of houses on the map were kept constant across all participants and conditions. House positions were rotated 45°clockwise about the centre of the map for the rotated condition (see Figure 3.2).

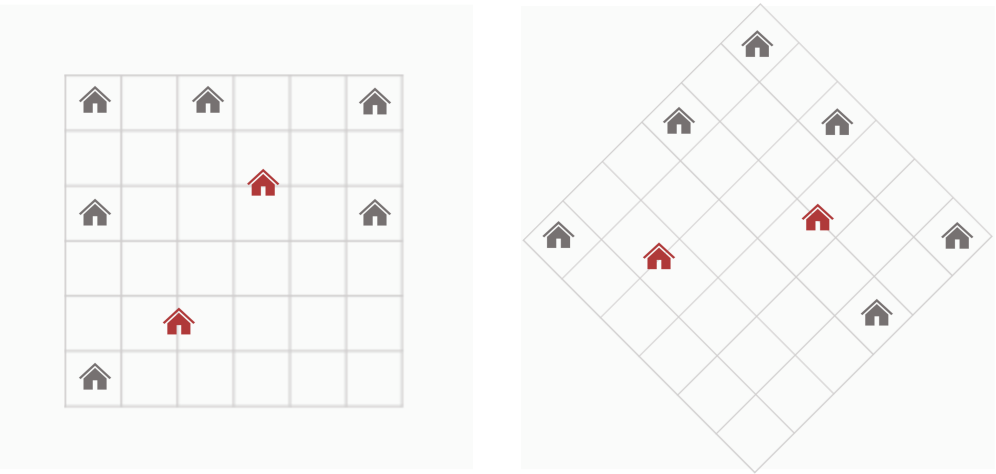


Figure 3.2: Neighbourhood maps used in the PAL task. The unrotated map is on the left, and rotated on the right. Participants were assigned to a rotation condition at the beginning of the experiment, and learned where each monster lived on their assigned map through paired-associate learning. The map of grey houses was visible to participants throughout the experiment. The red houses were only shown at the end of the experiment, to evaluate zero-shot generalisation based on systems alignment. Grid lines are shown for reference only, and were not visible during the experiment.

3.1.3 Monster Stimuli

Stimuli were generated in the free and open-source graphics editor Krita (V 4.2.2). All monsters had identical body shapes, hair, arms and legs. The monsters' eyes were orientation gratings.

Eye orientation took values between 5°and 85°from the horizontal, and body colour took values along a perceptually uniform trajectory from from blue to green. Feature spaces were generated such that the mapping of stimulus features onto spatial dimensions was randomised by participant. The direction

of variation along each spatial axis (e.g whether a green or blue monster was at the top of the map when colour was mapped onto the Y-axis) was also randomised independently for each feature dimension. This yielded a total of 8 possible feature spaces, all with the same range of values for each feature.

Stimulus sets for each participant were constructed based on the aligned condition in their randomly assigned feature space: stimuli were selected from positions in the 2D feature space which corresponded with the house positions in Figure 3.2. For participants in the misaligned condition, the stimuli in this constructed set were randomly assigned to houses in the neighbourhood.

Eye orientation

Sinusoidal orientation gratings (or Gabor patches) with a fixed spatial frequency of 5Hz were used as the monsters' eyes. The minimum rotation from horizontal was 5°, and the maximum was 85°. Prior studies have demonstrated that just noticeable differences (JND) in orientation are smaller than 1°(Vogels and Orban, 1985). The minimum difference between Gabor patch angles sampled for our stimuli was 32°for main trial stimuli and 8°for generalisation stimuli.

Body colour

This study required that stimuli could be generated at perceptually uniform intervals in the colour dimension, and that the colour values for neighbouring stimuli were perceptually distinct. To meet these criteria, we sampled colours along a linear trajectory in CIECAM02 Uniform Colour Space (CAM02-UCS) (Moroney et al., 2002). CAM02-UCS is a state-of-the-art uniform colour space, which outperforms previous spaces in modelling perceptual distances (Luo et al., 2006). The linear path in CAM02-UCS and corresponding colour scheme were generated using the `viscm` tool (Van der Walt and Smith, 2015, July 6–12).

For the main trials, we took 6 equally spaced values from this linear trajectory in CAM02-UCS. The CAM02-UCS and its predecessors were designed

such that 1 unit distance in the space corresponds to a JND in perception (Mokrzycki and Tatol, 2011). Kuehni (2016) investigated the relationship between JND in colour and the distances in CIECAM02-UCS experimentally, finding that 0.5 units in CAM02-UCS on average corresponded to a JND. Luo et al. (2006) demonstrates colour difference perceptibility in CAM02-UCS by plotting chromatic discrimination ellipses in the space, demonstrating that no difference thresholds perception distances in this space exceed 5 (Luo and Rigg, 1986; Melgosa et al., 1997). The ΔE between our colours in CAM02-UCS, calculated as the Euclidean distance in the space (Luo et al., 2006), is 12.3 - greater than even the most conservative JND values.

3.1.4 Procedure

The experiment was composed of three phases: pre-exposure, paired-associate learning and generalisation. In the pre-exposure phase, participants were familiarised with the monster feature space. In the paired-associate learning phase, participants learned to associate monsters with their homes over a series of trials. In the generalisation phase, participants' ability to generalise their learning to a new monster was tested. Each phase is elaborated upon below.

Pre-exposure

Before the task began, participants were given instructions for the task and were pre-exposed to the full set of stimuli. They were shown a pair of gifs, which cycled through the full range of feature values for monster colour and eye orientation respectively. The text on this page drew participant attention to the two dimensions of variation in the stimuli.

Paired-associate learning (PAL) task

The task procedure consisted of two trial types: *active* trials, in which participants were presented with a monster and asked to click on the home in which

they thought it lived, and *passive* trials, in which participants were shown the monsters' correct homes one by one. Trials were presented in separate blocks of active trials and passive trials, wherein every block contained one trial for each of the six stimuli in the set. The order of stimuli was randomised within each block. Each of the five active blocks was preceded by a set of two blocks of passive trials.

Prior to each set of passive blocks, participants landed on a break screen which prompted them to click a 'Continue' button to play the passive blocks. In each passive trial, the home whose resident was about to be revealed was cued with a grey border for 1 second. The resident monster was then shown in the home for 3 seconds before disappearing. The next home was cued after a 1 second break.

After two blocks of passive trials, participants moved on to a block of active trials. An example active trial screen is shown in Figure 3.3. On each trial, one monster was shown in the 'Holding Pen' on the left of the screen. Participants were instructed to click on the house on the map in which they thought the monster in the holding pen belonged. They could amend their choice as desired, and all clicks were recorded. Participants were instructed to click the 'Submit' button on the right hand side of the screen once they were happy with their choice.

The remaining five stimuli in the set were visible in a grid under the heading 'Other monsters' in the bottom left-hand corner of the screen. The arrangement of these stimuli was randomised, and was re-shuffled on each page load (i.e when the next trial was loaded or participants refreshed the page). This eliminated the possibility that participants could learn a mapping between between grid and map locations.

After submitting their response, participants recieved feedback. Once a participant submitted their response for a trial, a feedback screen was displayed for three seconds. This screen indicated whether their response had been correct or incorrect. If correct, participants advanced to the next trial

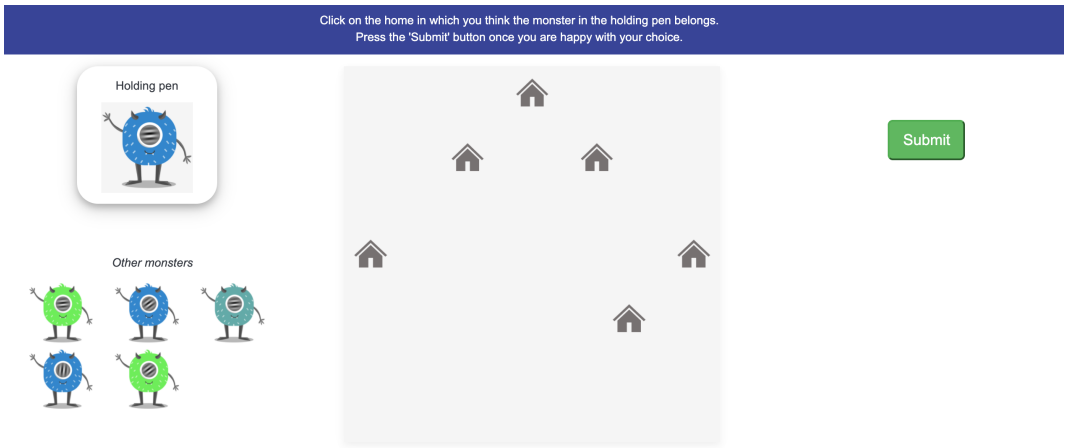


Figure 3.3: Example of an active trial screen in the rotated condition

automatically after three seconds. If incorrect, participants were prompted to click on the correct home which was highlighted with a grey box. Once they had clicked the correct home, they advanced to the next trial.

Generalisation task

After the PAL task was complete, participants were told that a new monster had moved to the neighbourhood, and that they were going to choose where they thought it should live on the map. The new monster was shown in the holding pen and there were two new homes to choose from in the map locations indicated by red houses in Figure 3.2. The monster's colour and eye orientation were both as-yet unseen values.

The instructions stated that the homes that they had been using in the previous trials would be visible on the map, but were not options for the new monster as they were already occupied. The trial screen was almost identical to the PAL trial screen, but the 'Other monsters' grid was removed and the homes that were used for the PAL task were greyed out and unclickable. Participants clicked on their choice of home for the monster, and submitted their answer. They received no feedback for this trial, and were taken straight to the debriefing page.

In the aligned condition, the unseen monster's position within monster feature space corresponded to the position of one of the presented house options.

The monster-house pair was randomly selected from the two options for each participant.

3.1.5 Participants

$N = 491$ participants completed the experiment in total, all of whom were recruited via AMT. Participants were limited to residents of the US and Canada. We required participants to have completed ≥ 1000 prior tasks with an acceptance rate $\geq 95\%$. All participants provided their informed consent prior to participation, and the experiment complied with UCL's code of ethics. The task took approximately 15 minutes to complete, and participants were paid \$2.00 for their participation. One participant was excluded for submitting inaccurate responses in the demographic survey.

Identifying poor engagement If a participant was making an earnest attempt at the task, we would expect their responses to be distributed near-uniformly across the available house options. Participants whose responses were poorly distributed across the options might have repeatedly submitted the same house or alternated between a small number of houses, indicating poor engagement with the task. We sought to exclude poorly engaged participants from the analysis. Our exclusion criterion was based on the average *entropy* of a participant's responses across blocks, $\bar{H}_b$, which is maximised by a uniform distribution of responses across house options. We excluded the 10% of participants with the lowest $\bar{H}_b$.

For each participant on each block of trials b, we calculated the entropy of the response distribution across options using:

$$H_b = \sum_{i=1}^{6} \mathrm{P}(\mathrm{X}_i)\log_2(\mathrm{P}(\mathrm{X}_i)),$$

where $\mathrm{P}(\mathrm{X}_i)$ was the probability of the participant selecting house i in block b, calculated as $\mathrm{P}(\mathrm{X}_i) = \frac{\sum_{t=1}^{6}[\mathrm{X}_t=i]}{6}$ for trial $t = (1, .., 6)$ in block b. $\bar{H}_b$ for each participant was the mean of H_b taken across all the experimental blocks.

To assess $\bar{H}_b$ as a criterion for participant engagement, we examined the

relationship between $\bar{H}_b$ and performance in the final block of trials. If $\bar{H}_b$ were a sensible measure of engagement, we would expect a relationship between low values of $\bar{H}_b$ and poor performance on the final block of trials, indicating that participants whose responses were not evenly distributed across the space of house options were not learning the task as well as others. This investigation was performed blindly with respect to experimental condition. The plot in Figure 3.4 demonstrates that there is a strong positive correlation between $\bar{H}_b$ and accuracy in the final block of responses ($r_p = 0.753, p < .001$). Excluding the bottom 10% of participants yielded an exclusion threshold $\bar{H}_b < 0.131$, visualised in Figure 3.4.

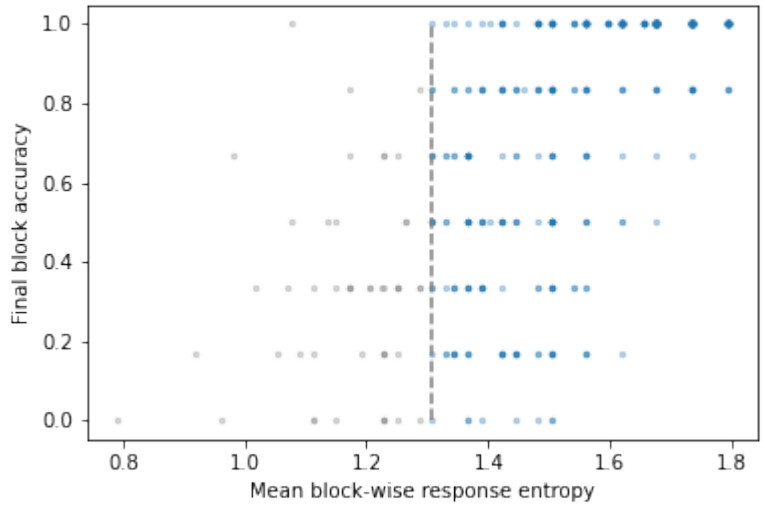


Figure 3.4: Relationship between final block accuracy and mean block-wise response entropy for all participants. Grey points represent excluded participants; blue points represent remaining sample after entropy threshold is applied.

The distribution of participants across conditions pre- and post-application of the entropy threshold is shown in Table 3.1. A χ^2 test comparing the proportions of participants by condition in the pre- and post-criterion samples reveals no significant difference in the impact of the entropy filter between conditions ($\chi^2(3) = 0.335, p = .953$).

Resultant sample This resulted in $N = 443$ participants whose responses were included in the following analysis. The sample was 40.2% female, with ages ranging from 20 to 72 ($\bar{x} = 45.5$ years, $\sigma = 14.7$ years). Following the random assignment of participants to conditions, a one-way ANOVA finds

		Pre-criterion	Post-criterion
Aligned	Unrotated	123 (.250)	110 (.248)
	Rotated	124 (.252)	112 (.251)
Misaligned	Unrotated	121 (.246)	106 (.239)
	Rotated	123 (.250)	115 (.262)
	Total N	491	443

Table 3.1: Distribution of participants across conditions pre- and post-application of the entropy-based exclusion criterion. Proportions of each condition in the total pre- and post-criterion samples respectively are shown in parentheses.

no significant difference in participant ages between conditions ($F(3, 439) = 0.487, p = .692$). A χ^2 test also finds no significant difference in the proportions of females between conditions ($\chi^2(3) = 2.42, p = .491$).

3.2 Experimental Results

3.2.1 Paired-associate learning

To evaluate how each condition impacts learning we examine two different measures: *response accuracy* and *distance error*. Response accuracy measures if a participant correctly mapped a monster. Distance error measures the distance between the chosen home and the correct home. If the monster is placed in the correct home, response accuracy is 1 and distance error is 0.

Analyses revealed significant main effects of alignment condition on both response accuracy and distance error. Results for block-wise means of response accuracy and distance error are shown in Figure 3.5. These results support our hypotheses that (i) learning is more successful in the PAL task when spaces are alignable than when they are not, and (ii) participants learning the PAL task across an alignable pair of spaces place the monster in homes with smaller distance error than participants learning paired associates in non-alignable spaces.

Results for both dependent variables were analysed using mixed-design ANOVAs. In each case, block was included as a within-subjects factor, and alignment and rotation conditions were included as between-subjects factors. Analyses were conducted using the package `ez` in R (Lawrence and Lawrence,

2016).

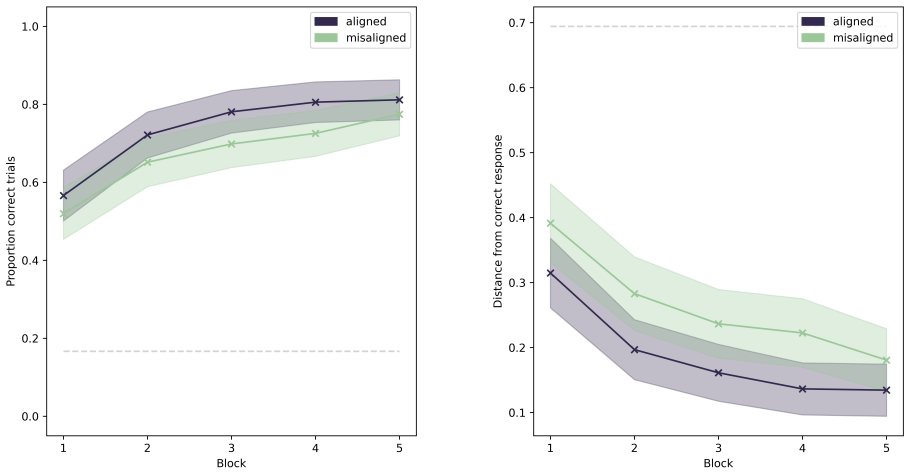


Figure 3.5: Results by alignment condition for (a) mean response accuracy and (b) mean distance error by experiment block. Blue lines show mean performance for participants in the aligned condition; red lines show mean performance for participants in the misaligned condition. Shaded areas show the 95% CI about group means.

In the ANOVA model fitted for block-wise mean response accuracy, Mauchly's test of sphericity indicated a violation of the sphericity assumption, therefore the Huynh-Feldt correction ($\epsilon = 0.83$) was used to appropriately adjust the degrees of freedom. Significant main effects of alignment condition ($F(1, 425) = 5.70, p = .017$) and block ($F(3.31, 1408.84) = 135.83, p < .001$) were found, but not of rotation condition ($F(1, 425) = 0.62, p = .430$) nor of any interaction terms. Full results for the ANOVA are shown in Table 3.2.

Predictor	df	ϵ	F	p
Alignment condition	$(1, 425)$		5.70	.017 *
Rotation condition	$(1, 425)$		0.62	.430
Alignment x Rotation	$(1, 425)$		0.17	.678
Block	$(3.31, 1408.84)$	0.83	135.83	$< .001$ *
Alignment x Block	$(3.31, 1408.84)$	0.83	1.23	.297
Rotation x Block	$(3.31, 1408.84)$	0.83	0.99	.402
Alignment x Rotation x Block	$(3.31, 1408.84)$	0.83	1.98	.108

Table 3.2: Results for repeated-measures ANOVA for block-wise mean accuracy. df = degrees of freedom; ϵ = Huynh-Feldt correction factor for violation of sphericity assumption.

The ANOVA model for block-wise mean distance error also violated the sphericity assumption according to Mauchly's test of sphericity. Degrees of freedom were adjusted accordingly using the Huynh-Feldt correction ($\epsilon = 0.82$). Significant main effects of alignment condition ($F(1, 425) = 13.67, p =<$

.001) and block ($F(3.29, 1398.98) = 120.16, p < .001$) were found, but not of rotation condition ($F(1, 425) = 0.62, p = .430$) nor of any interaction terms. Full results for the ANOVA are shown in Table 3.3.

Predictor	df	ϵ	F	p
Alignment condition	$(1, 425)$		13.67	$< .001$ *
Rotation condition	$(1, 425)$		0.62	.430
Alignment x Rotation	$(1, 425)$		0.08	.779
Block	$(3.29, 1398.98)$	0.82	120.16	$< .001$ *
Alignment x Block	$(3.29, 1398.98)$	0.82	1.21	.303
Rotation x Block	$(3.29, 1398.98)$	0.82	0.67	.584
Alignment x Rotation x Block	$(3.29, 1398.98)$	0.82	2.19	.081

Table 3.3: Results for repeated-measures ANOVA for block-wise mean distance error. df = degrees of freedom; ϵ = Huynh-Feldt correction factor for violation of sphericity assumption.

Considering the results visualised in Figure 3.5, it is worth noting that participants in the misaligned condition take all 5 blocks of trials to perform at the same standard reached in block 2 by those in the aligned condition - that is more than double the number of trials.

3.2.2 Generalisation

Our findings in the zero-shot learning trial support the notion that mapping between alignable systems enables generalisation to unseen examples.

Generalisation analyses are performed on the aligned condition only. Results for the misaligned participants were statistically indistinguishable from chance. This is as expected, given that there was no meaningful 'correct' response for these participants.

In the generalisation trial, 131 of the 222 participants in the aligned condition (59.0%) selected the correct house for the unseen monster, according to its position within an alignable system. This result is significantly above chance for $\alpha = 0.05$ ($\chi^2(1) = 7.21, p = .007$), supporting the hypothesis that participants who learn to align across are able to generalise to unseen mappings between the alignable structures.

A χ^2 test found no significant difference between rotated and unrotated conditions in generalisation accuracy where the systems were alignable ($\chi^2(1) =$

$0.03, p = .874$). This provides no support for the hypothesis that the ability to generalise across aligned systems depends on privileged sets of axes.

3.3 Modelling

There are range of cognitive strategies participants may use to complete the PAL task, each of which motivates a model in this portion of the study. We identify the best-fit model type for each participant, and compare the winning model counts within aligned and misaligned learning conditions. This allows us to better understand the distributions of learning strategies used in each. The strategy and implementation of each model is summarised below.

Classifier The Classifier model simulates a blind memorisation strategy, which makes no use of the 2D space. This strategy treats stimuli as unrelated from one another, and simply rote-learns an associated house for each monster. The Classifier is a multilayer perceptron (MLP) that takes as input a monster's feature coordinates and outputs a categorical prediction corresponding to a particular house.

The Classifier is comprised of an input layer, ReLU activation function, one fully-connected hidden layer of size 100 and output layer of size 6, corresponding to the $n = 6$ homes in which a stimulus could be placed on each trial. The input to the Classifier was the 2D coordinate vector of the stimulus in feature space, $\boldsymbol{x}$, normalised such that $x_d \in (0, 1)$ for $d \in \{1, 2\}$. The output vector was fed into a softmax function with temperature parameter T to produce a probability distribution across classes.

Regression The Regression model simulates a strategy which maps monsters into the 2D space of the neighbourhood, demonstrating an appreciation of the continuous nature of the feature space. The Regression model is a MLP that takes as input a monster's feature coordinates and outputs the coordinates of the correct house.

The Regression model $F(.)$ is comprised of an input layer, ReLU activation function, one fully-connected layer of size 100 and output layer of size 2. The

input to the model was the coordinate vector of the stimulus in feature space, $\boldsymbol{x}$, normalised such that $x_d \in (0, 1)$ for $d \in \{1, 2\}$. A sigmoid activation function was applied to model outputs to constrain output values such that $y_d \in (0, 1)$ for $d \in \{1, 2\}$. In other words, the MLP performed a mapping $F : X \to Y$ from stimulus space X to house space Y. To generate a probability distribution across house options, the Euclidean distance between the model output and each house option was subtracted from $\sqrt{2}$ (the maximum distance between points in the normalised space), yielding a measure of similarity which took values in range $(0, \sqrt{2})$. If the model had mapped a stimulus perfectly onto a house, this transformation would return its maximum value of $\sqrt{2}$, and conversely if a stimulus was mapped as far as possible from a house the value would be 0. The resultant distributions were fed into a softmax function with temperature parameter T to generate a probability distribution across houses for the stimulus according to the model.

Regression + Aligner model The Regression + Aligner model also maps monsters into the neighbourhood, with an added assumption that the systems of houses and monsters should be aligned. On each trial, it updates its internal representions based on the trial feedback together with its knowledge of the structural relationships within systems.

The Regression + Aligner model has a similar structure to the Regression model, with some key additions corresponding to the putative 'space alignment' mechanism. Firstly, while the Regression model consisted of one MLP $F(.)$, which mapped from stimuli $\boldsymbol{X}$ to houses $\boldsymbol{Y}$, the Regression + Aligner model consisted of two MLPs: $F(.)$ and $G(.)$. These perform mappings $F : X \to Y$ and $G : Y \to X$ respectively. This is visualised in the leftmost panel of Figure 3.6.

Secondly, the aligner minimised two additional unsupervised loss components: a *cycle consistency loss* term ($\mathcal{L}_{cyc}$) and a *distribution loss* term ($\mathcal{L}_{dist}$). Inspired by the work of Zhu et al. (2017), $\mathcal{L}_{cyc}$ is defined as the mean Euclidean distance between input stimuli $\boldsymbol{X}$ and the recovered estimates $\boldsymbol{X}''$, generated

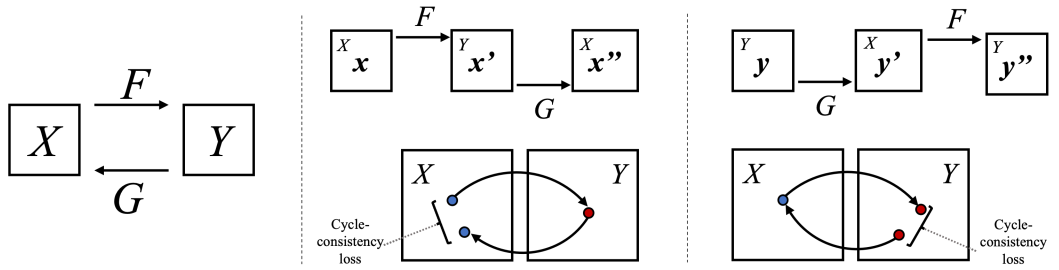


Figure 3.6: Illustration of cycle consistency loss $\mathcal{L}_{cyc}$, adapted from Zhu et al. (2017). The Aligner model is comprised of two MLPs $F(.)$ and $G(.)$, visualised in the leftmost panel. $\mathcal{L}_{cyc}$ measures the average distance between each point in its original space, $\boldsymbol{x}$, and its reconstruction in the same space $\boldsymbol{x}''$ generated by the mapping $\boldsymbol{x}'' = F(G(\boldsymbol{x}))$

by mapping via both MLPs: $\boldsymbol{X}'' = G(F(\boldsymbol{X}))$. This is visualised in Figure 3.6.

$\mathcal{L}_{dist}$ is visualised in Figure 3.7. In space Y, it is defined as the mean negative log likelihood (NLL) of all $F(\boldsymbol{X})$ as samples from a Gaussian mixture model comprised of 2D Gaussian kernels placed on Y. $\mathcal{L}_{dist}$ is minimised when all $F(\boldsymbol{X})$ are mapped directly onto Y. Further details of both loss terms are provided in B1.

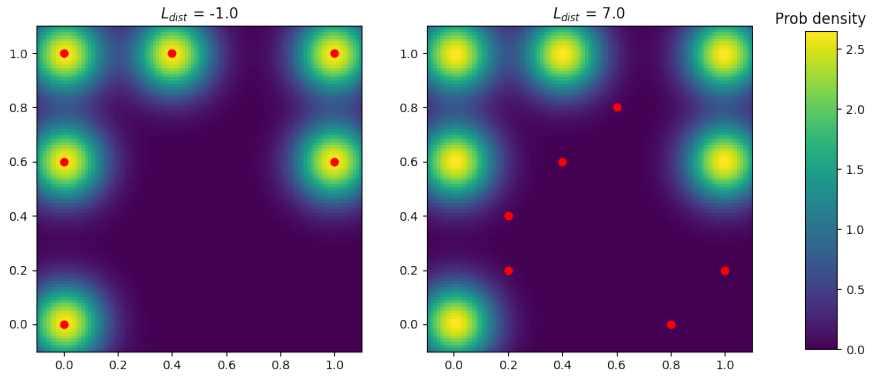


Figure 3.7: Visualisation of distribution loss $\mathcal{L}_{dist}$ for a low (left) and high (right) loss mapping. Red points represent $\boldsymbol{X}'$, overlaid on a heatmap representing the probability density of GMM_Y.

3.3.1 Training and model fitting

For every participant, each model type was fitted to see how well it could replicate the participant's behaviour in the PAL task. Each model's hyperparameters were selected to minimise the total negative log likelihood (NLL) of the participant's submitted responses across all trials. Both in hyperparameter

selection and in final training, the sequence of inputs in model training was matched to the sequence seen by the relevant participant in the experiment. This stimulus sequence included both active and passive trials. Passive trials were masked from the NLL calculation in hyperparameter optimisation.

A random model was included as a control. This model selected a random home for the presented stimulus on each trial. It did not learn, and no hyperparameters were fitted to individual participant data.

For each participant, the best fitting model type was that which minimised the Akaike Information Criterion (AIC), where $AIC = 2k + 2NLL$. k is the number of hyperparameters which could be varied to fit the participant data for each model type.

All models were built and trained using `pytorch`. Model weights in all cases were initialised with Xavier uniform initialisation. On each trial, models performed 30 update steps using stochastic gradient descent (SGD) with constant lr. Multiple steps were required to balance the need for fast learning (owing to the small number of trials) with the instability of high learning rates. Preliminary tests found that 10 gradient steps per trial was the maximum value required for any model to reach optimal performance.

To prevent any probabilities from reaching zero and causing computational issues, we took the maximum of each resultant probability and a small ϵ ($\epsilon = 10^{-30}$), and re-normalised the distribution. In model training, the loss term on each trial was the negative log-likelihood (NLL) of the correct response according to this distribution.

Hyperparameter optimisation was performed using the `hyperopt` package in python. Optimisation was performed over 150 evaluations for each model of each participant, using the Tree Parzen Estimator (TPE) method. Preliminary testing found that the success of the Classifier model in learning the task was particularly sensitive to initialisation, while the Regression and Regression + Aligner models were more stable. As such, the classifier was trained three times with each set of hyperparameters tested, and the minimum NLL across

the three iterations was taken as the score for those hyperparameters.

In all three models described above, the softmax temperature parameter T and learning rate lr were hyperparameters. One final hyperparameter, α, described each participant's probability of choosing according to the model on any given trial. The probability of choosing a random house was therefore $(1 - \alpha)$. Where random variable Y_t is the model's house choice on trial t, the probability of a participant choosing house y on trial t was modelled as: $P(y) = \alpha P(Y_t = y) + (1 - \alpha)(\frac{1}{6})$

The Regression + Aligner model had two additional hyperparameters, λ_{cyc} and λ_{dist}. This yielded a total number of hyperparameters $k = 3$ for the Classifier and Regression models, and $k = 5$ for the Regression + Aligner model.

3.3.2 Modelling results

Model fitting finds that the majority of participants in the aligned condition are best fitted by the Regression + Aligner model. This supports our hypothesis that participant responses in the aligned condition would be best captured by a model with an alignment mechanism. Surprisingly, we also find that the majority of participants in the misaligned condition are best fitted by the Regression + Aligner model. Models were fitted to minimise the NLL of participants' responses in the experiment, not to best predict the correct answer, so this result suggests that participants may attempt alignment by default in learning, and seek out alignable signals even when this leads to errors.

The left panel of Figure 3.8 shows counts of the best fitting model types within each condition. The Regression + Aligner model was the best fitting model for the majority of participants in both aligned (84.2%, $\chi^2(3) = 418.40$, $p < .001$) and misaligned (54.2%, $\chi^2(3) = 123.16$, $p < .001$) conditions.

The AICs for best-fit models of each type relative to the random model are shown in the right-hand panel of Figure 3.8. Note that among those best

fitted by the Regression + Aligner model, improvement over random for aligned participants is far greater than for misaligned participants. This is consistent with the fact that the alignment signal was misleading for participants in the misaligned condition. The result that misaligned participants are best-fitted by the Regression + Aligner model means it was the most predictive model of their errors, but misaligned participants would have to overcome the tendency to seek out alignment in order to respond correctly in the task, leading to poorer model fit.

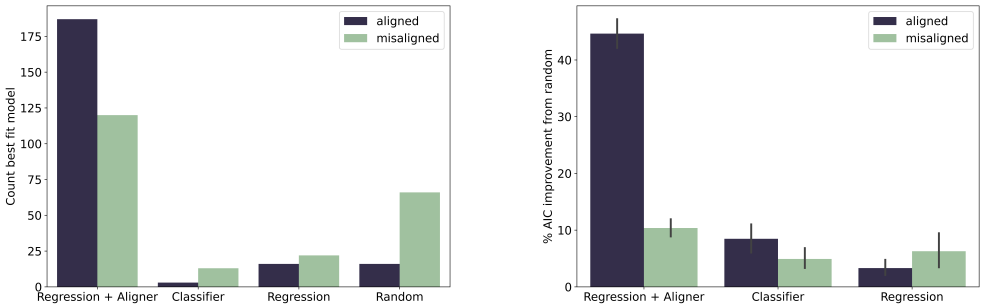


Figure 3.8: Best fitting models by participant. Left: Frequencies of each model type being chosen as best fitting model on the basis of AIC; right: improvement of each fitted model over random on AIC.

3.4 Discussion

This thesis' overarching aim is to explore the value of systems alignment for learning in naturalistic contexts. The previous chapter used simulations based on naturalistic data to demonstrate that alignment signals are valuable for early concept learning in the real world. The current chapter moved this investigation into a lab setting, to test whether and how humans benefit from these alignment-based signals where they exist. To understand how systems alignment operates in conjunction with the other signals available in real-world learning scenarios, the impact of alignment on learning was explored in the presence of supervisory signals.

This chapter has developed the contribution of this thesis in two ways: first, the behavioural experiment provides evidence that humans benefit from alignability when learning to map between spaces, both in terms of the effi-

ciency of learning and the ability to accurately generalise to previously unseen examples. Secondly, behavioural modelling results demonstrate that an alignment mechanism is well-placed to account for how humans learn the paired associate learning task relative to traditional models.

The experimental results suggest that aligned spaces facilitate more efficient cross-system learning than misaligned spaces. In the context of Roads and Love (2020)'s finding that spaces derived from unimodal distributional semantics are alignable across modalities, this suggests that systems alignment could support cross-modal learning in the real-world. Our significant result for the generalisation task suggests that alignable spaces could facilitate asynchronous integration of multi-modal information in human concept learning (Fourtassi and Dupoux, 2016; Samuelson et al., 2011; Socher et al., 2013). Future work could explore how alignment applies to different domains and types of similarity relationship. With Socher et al. (2013)'s computational work in mind, which demonstrates that zero-shot learning of multimodal concepts is possible by transferring information between unimodal distributions, our significant result for the generalisation trial suggests that alignable spaces could support the asynchronous integration of multi-modal information in concept learning (Fourtassi and Dupoux, 2016; Samuelson et al., 2011).

The results of the model-fitting provide evidence in favour of the hypothesis that an alignment process may be recruited when learning to map between systems. Models which included an unsupervised loss term for whole-system alignment were superior on AIC for the majority of participants. In the context of indeterminacy of reference (Quine, 1960) and often infrequent supervised learning episodes (Lieven, 1994), the incremental benefit of an unsupervised aligner loss term suggests a place for an alignment mechanism in explanations of humans' concept acquisition in the real world.

The relative success of the Regression + Aligner model in fitting participant responses, even in the misaligned condition, suggests that participants attempt to align systems even where they are not alignable, and make errors consistent

with this approach. So while our original hypothesis only pertained to the aligned condition, the fact that the majority of participants in the misaligned condition are also best modelled by the Regression + Aligner provides further evidence for alignment mechanisms in learning.

In the context of concept learning, systems alignment mechanisms could provide an account of how amodal concept representations incorporate information from different modalities (Patterson et al., 2007b; Ralph et al., 2017b; Popham et al., 2021).

Here, we have explored the role of alignment signals in supervised learning. Future work may seek to understand how people use these signals in more ecological multimodal learning contexts, where learning signals are noisier. Cross-situational learning, for example, provides participants with weak supervision across multiple training episodes (Smith and Yu, 2008; Yu and Smith, 2007), and has been found to be enhanced by semantically themed encoding contexts (Chen and Yu, 2017). An examination of the effect of alignable systems in a weakly-supervised context would further develop our understanding of how alignment signals are used in the real-world, and how these interact with other learning signals.

Of course, the scale of ecological alignment problems is much larger than those tested here, but the possibility remains that established learning processes are supplemented by these alignable signals. Indeed, larger systems have richer signals for alignment (Roads and Love, 2020; Goldstone and Rogosky, 2002). The relatively small effect size observed here may be attributed to the low difficulty of the task: with only 6 items to hold in memory, the task was intended to be learnable for the majority of participants even in the misaligned case. But the incremental benefit of cross-system alignment may increase with problem size, as the cognitive cost of aligning systems is overpowered by the cognitive cost of memorising individual mappings. A natural extension of this project could explore the role of alignment in learning for different problem sizes. Future work may seek to explore the role of alignable

spaces in ecological multimodal learning contexts, for example the mapping of concept labels to their referents. This is explored from a machine learning perspective in the next chapter.

In sum, our findings provide evidence for the value of alignable spaces in accelerating human learning. Together with prior work demonstrating that real-world multimodal spaces are alignable, this opens an avenue of exploration regarding how humans may tackle referential ambiguity in concept learning, and how we learn from the statistics of our noisy environments more broadly.

The modelling portion of this chapter demonstrated that an asynchronous, unsupervised alignment mechanism was a strong candidate explanation of how people use this information to support learning. The next chapter of this thesis extends the investigation of alignment mechanisms, applying alignment to larger-scale problems with data from naturalistic environments. Applying these human behavioural findings to computational approaches, we also explore the use of alignment mechanisms in cross-modal machine learning problems.

Chapter 4

Modelling alignability at scale

4.1 Introduction

In Chapter 3, it was shown that humans benefit from alignment when learning to map between systems. Computational modelling of this effect found that an asynchronous alignment mechanism was the best model for capturing how humans learned to map across systems. In this chapter, I evaluate whether machine learning algorithms can take inspiration from this process, and capitalise on alignment signals to facilitate cross-system learning from naturalistic data.

Prior work demonstrating that real-world cross-modal systems possess shared underlying structure at scale (Roads and Love, 2020) suggests that alignment mechanisms could prove valuable in cross-modal machine learning. Conservatively, cross-modal alignment signals may provide a useful prior, which could improve the efficiency of learning and inject a valuable signal in instances where training data is restricted (Zaadnoordijk et al., 2022). This perspective is further supported by the aforementioned behavioural experiment discussed in Chapter 3 of this thesis, which indicated that humans make use of alignable spaces in learning environments even when supervision signals are available. Taking the implications of alignment further, alignable signals could be sufficient to perform fully unsupervised alignment across modalities, allowing mapping between visual objects and words (for example) to be learned with

no supervised examples.

In the real world, there is a huge deal more data to align than the 6 items mapped in the study presented in Chapter 3, and structural correspondences are undoubtedly noisier than this controlled example. While Roads and Love (2020) demonstrated that the alignment between inter-concept relationships across systems provides a signal for mapping accuracy, there has to date been no in-depth exploration of how these alignment signals could be used to find cross-modal mappings at scale. This chapter uses simulations based on real-world embedding data to compare scalable algorithms for cross-modal alignment.

The analyses in this chapter are split into two main sections. First, I tested a range of loss function optimisations and candidate alignment algorithms on the unsupervised alignment problem. Using a combination of synthetic examples and naturalistic tests, I do not find a viable solution for completely unsupervised cross-modal alignment at scale. However, in line with prior work, both in Chapter 2 of this thesis and elsewhere (Akata et al., 2015; Socher et al., 2013; Frome et al., 2013), I find that alignment yields promising solutions for novel concepts when some concepts are known.

In the second section of the chapter, I use these promising alignment methods to generate priors across classes in an image classification task. The prior is generated based on knowledge of classes which are not included in the classification problem. To foreshadow results, I find that an alignment mechanism with an unsupervised cycle loss component yields the strongest prior. Interestingly, this is only true when the prior is trained without full supervision.

4.1.1 Relevant prior work

While there are multiple related streams of prior work, there is no existing model which attempts the unsupervised alignment of information across multiple naturalistic modalities.

Machine learning approaches to finding mappings between modalities have

largely been supervised or semi-supervised. Supervised approaches include those presented by Socher et al. (2013), where some image-word mappings are used to learn a mapping between spaces, which is then used to learn mappings for novel items. Taking a similar approach, Frome et al. (2013)'s DeViSE and Akata et al. (2015)'s Structured Joint Embeddings train joint linguistic and visual embeddings by training an embedding model to map visual representations onto their known labels in image space.

Unsupervised mappings between conceptual systems based on similarity structures have been discussed by Goldstone and Rogosky (2002), where a constraint satisfaction network was used to map between concepts based on idiosyncratic similarity judgments. In machine translation contexts, where different linguistic representations are mapped onto eachother, success in unsupervised mapping has been achieved by using adversarial networks (Zhang et al., 2017; Conneau et al., 2017). Unsupervised models in cross-modal tasks largely take the form of generative adversarial networks (GANs), generally applied to problems of a different nature, such as image captioning and text-to-image synthesis (Reed et al., 2016; Shetty et al., 2017; Dai et al., 2017; Gu et al., 2019; Feng et al., 2019).

Optimal transport methods have also been used for structural alignment problems, particularly with applications to graph matching (Titouan et al., 2019; Seguy et al., 2017). Optimal transport methods optimise distribution matches by learning to minimise the cost of transporting one distribution to another distribution. This has been successfully applied to image-text domain mapping problems, such as visual-question answering (Chen et al., 2020a).

Large-scale unsupervised learning problems can also be tackled using reinforcement-learning inspired techniques such as Monte-Carlo Tree Search, which has demonstrably been successful on tasks with large solution spaces (Browne et al., 2012; Silver et al., 2016, 2018; Pinheiro et al., 2016).

Models inspired by these related efforts are applied to the unsupervised alignment problem in this chapter.

4.1.2 Challenges for a cross-modal alignment algorithm

In order to guide the efforts to build an unsupervised alignment algorithm - the first focus of this chapter - let us consider some of the critical challenges that such an algorithm faces.

When it comes to learning cross-modal mappings at scale, the solution space quickly becomes intractably large for an exhaustive search of mappings. Increasing the number of concepts to be learned, N, vastly increases the number of possible mappings between systems, $N!$. For 10 concepts, the number of possible mappings is 3.6×10^6; for 50 concepts, this number rises to 3.0×10^{64}, and for 100, to 9.3×10^{157}.

Even in cases where the full set of mappings could be searched, unsupervised algorithms will also contend with the problem of misleading mappings. A misleading mapping is any incorrect mapping which has a higher score on the chosen objective function than the correct mapping does. In Roads and Love (2020), it was shown that the proportion of misleading mappings gets smaller as the number of concepts gets larger. However, together with the previous challenge of combinatorial explosion, this does not eliminate the problem of misleading mappings, as there may still be a large number of misleading mappings which cause problems for a candidate algorithm.

As discussed in Chapter 2 of this thesis, it is also possible that long-range similarity relationships do not show the same degree of cross-modal correspondence as close-range similarity relationships. While there is a meaningful answer to the question of whether a 'mouse' is more similar to a 'rat' or to a 'table', it is less meaningful to ask whether a 'mouse' is more similar to a 'glass' or to a 'shoe'. In line with this intuition, it was shown in Chapter 2 that long-range similarity relationships are less stable across multiple initialisations of an embedding than short-range relationships. Embeddings in different modalities can be considered as analogous to the multiple initialisations of the embedding algorithm tested in Chapter 2. As a result, while similarity structure is broadly reflected across modalities, noise may originate from arbitrary differences in

long-range inter-concept distances.

Furthermore, when working with high-dimensionsal embeddings (e.g, 50-dimensional word embeddings from GloVe), there is a risk of problems related to the curse of dimensionality (Bellman, 1966). For example, high-dimensionality can have unintuitive effects on inter-concept distance calculations (Aggarwal et al., 2001), which are the basis of alignment.

The following investigation aims to facilitate unsupervised alignment by addressing these challenges. A range of preprocessing steps (including dimensionality reduction and the transformation of similarity relationships) and modifications of the objective function are tested, to see if the number of misleading mappings can be minimised. Then, algorithms which have demonstrated potential to be robust to the large solution space are prioritised. The resultant testing framework is visualised in Figure 4.1

4.2 Unsupervised cross-modal alignment

4.2.1 Materials

Text embeddings

The pre-trained word embeddings used here were 50-dimensional GloVe text embeddings (Pennington et al., 2014). These embeddings are trained on 6 billion tokens from the Wikipedia2014 + GigaWord5 text corpus. The resultant vocabulary size is 400,000 tokens.

Image embeddings

The image embeddings used here are those used in Roads and Love (2020), derived by applying the GloVe algorithm (Pennington et al., 2014) to the Open Images V4 dataset (Kuznetsova et al., 2020). Open Images V4 is comprised of approximately 9.2 million images, all annotated to identify which of over 19,000 object classes they contain. Roads and Love (2020) construct a co-occurrence matrix by counting the images in which each object class co-occurs

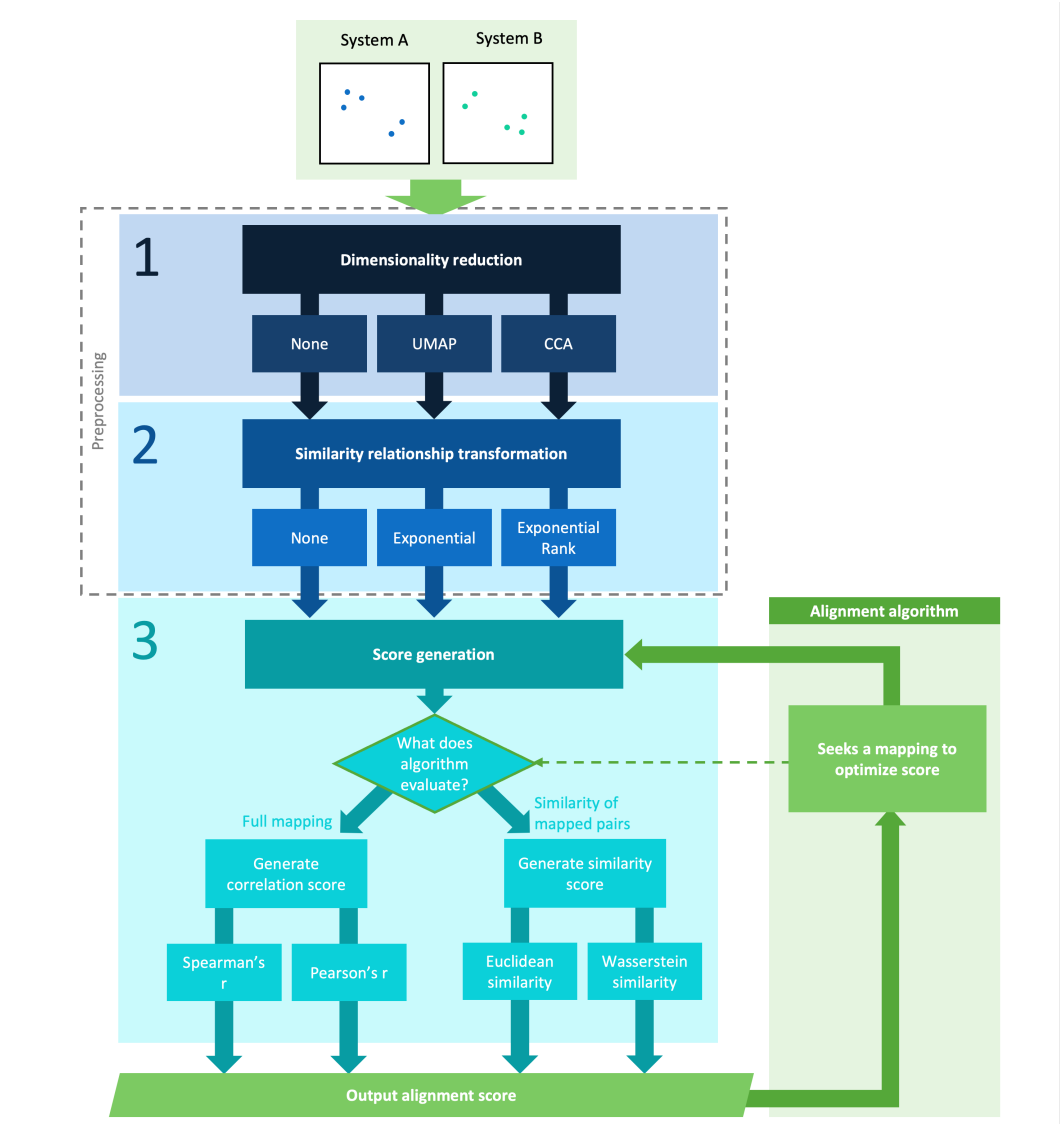


Figure 4.1: Figure showing the breakdown of alignment components which we explore in Chapter 4. Boxes 1 and 2 form the pre-processing steps which can be applied prior to score generation. Box 3 outlines the choices to be made regarding which function is used to generate the alignment score. Boxes 1, 2 and 3 together constitute the construction of the alignment score. The diagram also visualises how the score interacts with the alignment algorithm, which is explored in the second section of this chapter.

with each other class. This matrix is inputted to the GloVe algorithm, which generates the 10-dimensional image embeddings we use.

4.2.2 Optimising the objective for unsupervised alignment

The aim of this section is to identify an alignment scoring metric which is best positioned to yield success when used by an unsupervised alignment algorithm. We test a range of modifications to the alignment score, aiming to address the challenges of alignment discussed above and to incorporate insights from prior chapters.

The alignment scoring process is broken down into three key components: (1) transformations of the concept space (i.e., is dimensionality reduction applied); (2) the similarity relations which are compared across systems (i.e., are similarity relationships transformed within individual modalities) (3) the function for generating the score. The whole process is outlined in Figure 4.1.

The method used in step (3) is partially dictated by the alignment algorithm of choice. Some algorithms optimise a global score, which is calculated based on a complete proposed mapping. Others optimise a score which evaluates each pair of concepts mapped across systems, independent of how other items in the system are mapped.

Dimensionality reduction

To address concerns around the curse of dimensionality, raised in section 4.1.2 above, we explore methods of dimensionality reduction as a pre-processing step before score generation. The intuition here is that there may, for example, be certain dimensions of variation in the pretrained linguistic embedding which are redundant for the set of items (concrete nouns) being mapped to the visual domain. As such, when we look at the dimensions of variability of objects which occur in both visual and linguistic domains, we may find a smaller space of dimensions which is more relevant for alignment.

Dimensionality reduction is applied to the system representations, before the relationship between the alignment score and mapping accuracy is probed. Note that these methods are not themselves performing alignment - they are simply being explored in their ability to improve the success of alignment algorithms.

We test three dimensionality reduction techniques to explore their impact on the strength of alignment signals: Canonical Correlation Analysis (CCA), Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). Note that CCA requires a known mapping between systems in order for the space to be constructed, and as such it is not a viable candidate for use in the final unsupervised alignment algorithm. It is included only as a proof of concept and upper bound on dimensionality reduction performance. PCA and UMAP are performed on each space in isolation, and so are consistent with an unsupervised alignment algorithm.

Canonical correlation analysis (CCA) CCA was originally proposed by Hotelling (1992), as a method for finding a set of basis vectors to maximise the correlation between the projections of different sets of variables onto these basis vectors.

CCA acts as a supervised form of dimensionality reduction for both systems before alignment. It seeks a projection of each system into a common space, with a pre-specified number of dimensions k, where the common space maximises the correlation between matched items across its dimensions. By definition, this requires a known mapping between spaces. As such, in the terms of the original formulation in Hotelling (1992), dimensions in system X are considered one set of variables, and dimensions in system Y are considered the other. We test values of $k \in 2, 5, 10$. CCA was performed using the Python package `scikit-learn`.

Principal component analysis (PCA) PCA is the first of the unsupervised dimensionality techniques tested. Some unsupervised dimension reduction techniques focus on preserving global pairwise similarity structure

others prioritise local over global similarity structure preservation. PCA is an example of the former category, where global pairwise similarity structure is preserved. PCA derives a set of d orthogonal basis vectors, or *principal components*, onto which the d original dimensions of a set of points can be projected. They are ordered such that the first principal component captures the largest amount of variation possible in the data, the second principal component captures the second highest amount of variation, and so forth until the d^{th} principal component. Dimensionality reduction is performed by selecting only the first k principal components, where k is the desired number of dimensions in the transformed space and $k < d$. In our experiments, we test values of $k \in 2, 5, 10$. PCA was performed using the Python package `scikit-learn`.

Uniform Manifold Approximation and Projection (UMAP) By contrast, UMAP is an unsupervised dimensionality reduction method which focuses on the preservation of local similarity structure over global similarity structure. This class of algorithms are sometimes referred to as *neighbourhood based* dimensionality reduction (McInnes et al., 2018). In our analyses of early-acquired concepts in Chapter 2, we found that local structure was among the most important structural signals for promoting alignment, as local relationships were more likely to be preserved across multiple embedding initialisiations, and likely represented more stable semantic relationships. Thus, it was logical to test a dimensionality reduction algorithm which prioritised preserving these relationships, in the hopes of facilitating unsupervised alignment.

UMAP works by first constructing a weighted nearest neighbour graph for the points in the system. Each point in the system is a vertex V in the weighted graph. The number of nearest neighbours included, N_{neigh}, is a hyperparameter which determines the number of outgoing edges incident on each vertex. Edge weights are calculated based on the distance between vertices which are connected by an edge, with lower weights corresponding to higher distances. Weights are interpretable as the probability that the edge exists (McInnes et al., 2018).

A force directed graph layout algorithm is applied to the resultant weighted graph, where attractive forces along edges and repulsive forces between vertices are applied iteratively. The layout of the space is optimised with respect to these forces until a local minimum is reached. The number of target dimensions k and the minimum distance bewteen points in the final embedding space *minDist* are additional hyperparameters.

Values of $k \in 2, 5, 10$ and $N_{neigh} \in [2, 150]$ were tested. The default value of $minDist = 0.1$ in the python package `UMAP`, which was used to perform this dimensionality reduction, is retained.

Similarity relations

Raw inter-concept distance The raw inter-concept distance is the Euclidean distance between concepts in the relevant embedding space. Within each system, pairwise distances are scaled such that values fall within the range [0,1].

Exponential transform Based on the finding that longer-range similarity relationships are less stable across embedding initialisations (shown in Figure 2.10), there is reason to believe a new similarity function which down-weights the importance of long-range relationships could improve our capacity to identify appropriate mappings between systems.

An appropriate transformation of inter-concept distances would ensure that the contribution of differences between long-range relationships to the alignment score of two systems under a given mapping, was minimised. That is, if two items were deemed to be 'dissimilar' in one system and 'very dissimilar' in another, this would be negligably different from the case where they were deemed to be 'very dissimilar' in both systems (recall the intuitive example from section 4.1.2). Therefore, for an alignment algorithm to penalise a mapping on the basis of this kind of meaningless alignment may be counterproductive, as it steals attention in the correlation calculation from the more meaningful local connections.

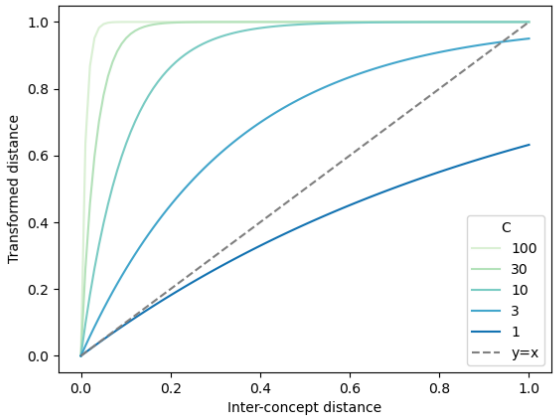


Figure 4.2: Example transformations of pairwise distances, to upweight differences in local relationships, according to $d_{trans} = 1 - e^{-Cd}$ for varying values of C. This is a visualisation of one of the preprocessing steps tested in an effort to identify the best way to calculate alignment score, such that it leads to accurate mappings.

Shepard (1987) proposes that human judgments of similarity follow an exponential law, and this has been shown to hold when tested on large volumes of naturalistic stimuli (Marjieh et al., 2023). Together with the finding that more proximal relationships are more valuable than distant ones for alignment, this leads us to explore exponentially transformed spaces for our naturalistic alignment problem. Examples of such a transformation are shown in Figure 4.2.

Exponential rank transform An alternative method transforming our alignment correlation function, is to apply a transformation to the ranked pairwise distance matrices, such that differences in smaller distances are upweighted and difference between larger distances are downweighted. The motivation here is that the distributions of inter-concept relationships are non-identical across systems (see Figure 4.3). By ranking pairwise distances within each system, our distance function can become agnostic to any relative skew in the distance distributions, potentially leading to a stronger signal for alignment.

First, we obtain matrices of the ranked inter-concept distances in systems X and Y, $\mathbf{R}_X$ and $\mathbf{R}_Y$ respectively. We then scale $\mathbf{R}_X$ and $\mathbf{R}_Y$ such that the maximum value of each is set to 1, by dividing by the maximum rank in each. Then, we can transform these values such that they satisfy our desired

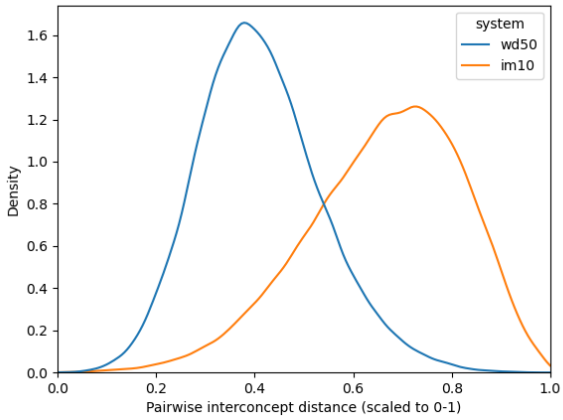


Figure 4.3: Non-identical distributions of pairwise distances across systems

transformation criteria, using: $R_{X,trans} = 1 - e^{-CR_X}$, where parameter C determines the relative weighting of local vs distant relations. Examples of this transformation for different values of C are shown in Figure 4.2.

To calculate the alignment score from these transformed ranks, we simply take the Pearson correlation of the transformed ranks. In this case, we use the same value of C in both modalities, as by applying ranks we remove any issues of distance distribution.

Scoring function

As previously mentioned, the options available for the scoring function depend on the scale at which alignment is being evaluated, which is in turn dictated by the alignment algorithm. This can be split into scores which evaluate (a) the alignment within a candidate pairwise mapping between two systems (full mapping score), or (b) the alignment of the similarity relationships between two individual items in different systems (pairwise score).

Visualised in Figure 4.4, the full mapping score gives a holistic alignment score, which measures alignment across the entire systems of relationships. The score is calculated by correlating the upper triangular portions of pairwise similarity matrices across systems, where the positions of columns and rows in the pairwise similarity matrices are determined by the candidate mapping of items across systems. In the orange box of Figure 4.4, for example, the

proposed mapping places item A from system X in correspondence with item 4 from system Y. As such, item A occupies the first row and column of system X's pairwise distance matrix, and item 4 occupies the first row and column of system Y's pairwise distance matrix. Item B in system X is mapped to item 2 in system Y, and so these items are both placed in the second position in the pairwise distance matrix, and so on. After the matrices are arranged in this way, the correlation between the upper triangular portions of the matrices quantifies the extent to which this mapping places the systems in alignment, by measuring the extent to which the structural relationships correspond across systems for mapped items. The correlation can either be the Spearman correlation, as was used in Roads and Love (2020), or the Pearson correlation. Rationale for these options is discussed below.

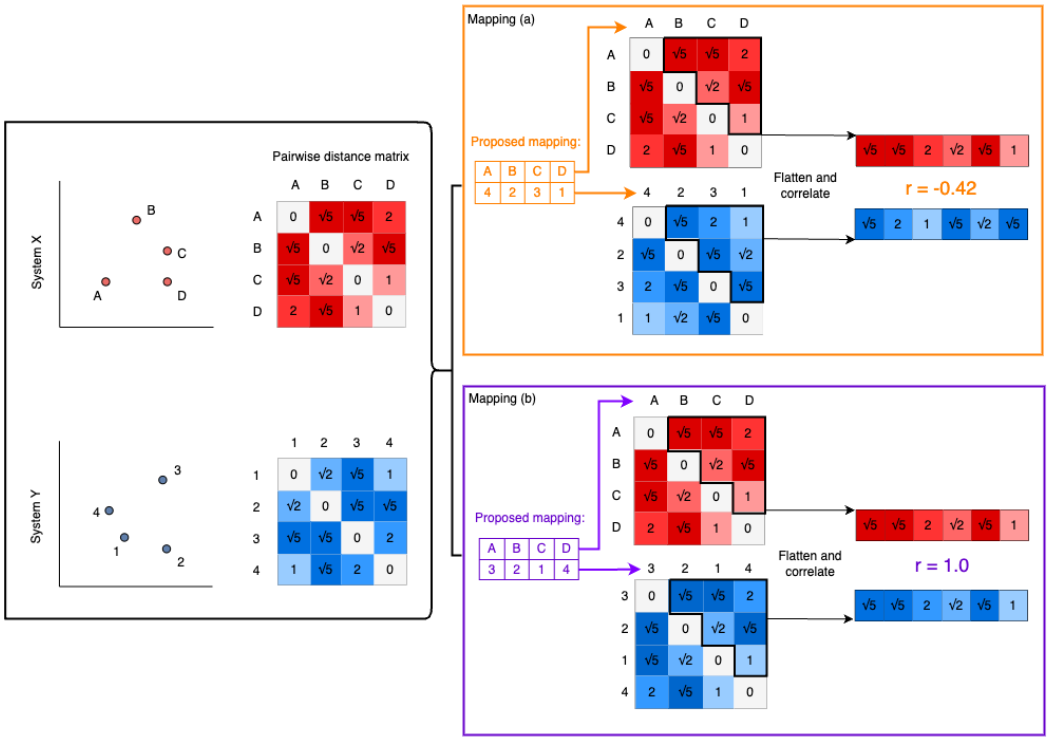


Figure 4.4: Example of full mapping score calculation, adapted from Roads and Love (2020). This is the calculation of the mapping score for algorithms which evaluate an entire proposed cross-system mapping.

The pairwise score is less holistic, and instead tries to adapt the principle of second-order isomorphism for use in algorithms which require a score to be evaluated for candidate pairings of individual items across systems. In this case, the score is calculated by (1) computing the similarity relationships within each system, (2) sorting the pairwise distances for each item, to ob-

tain a 'fingerprint' for each item in each system, (3) obtaining the pairwise distances between the two systems' fingerprint matrices. These distances can be calculated either as a Euclidean or a Wasserstein distance, both of which are explained below. The output of this is a pairwise cost matrix, which gives the cost of placing each pair of items in correspondence across systems. A cross-system mapping can then be obtained by optimising based on this cost matrix.

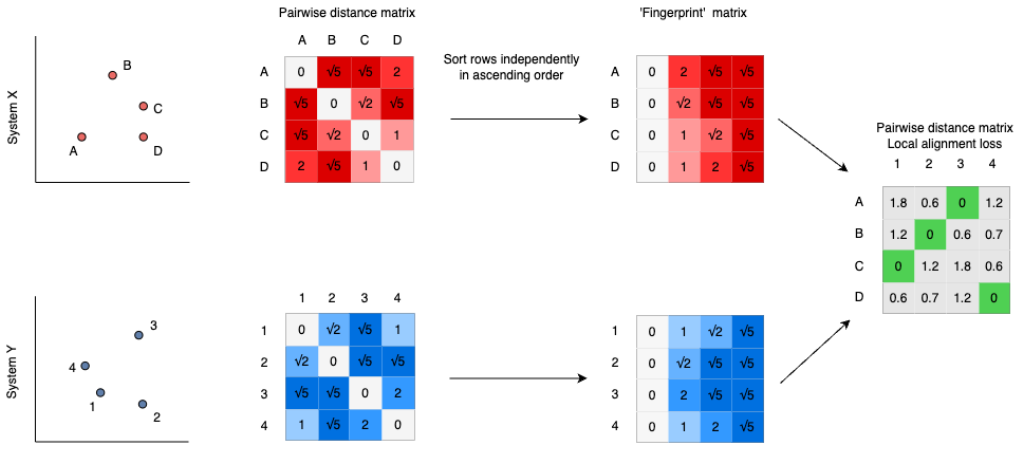


Figure 4.5: Example of pairwise cost calculation. This is the calculation used for algorithms which evaluate cross-system mappings for individual points. The 'local alignment loss' matrix outputted by the calculation can be inputted as a cost matrix to an algorithm which makes pairs to minimise global cost.

The next sections discuss variations within each score type (full mapping score and pairwise score) which are to be compared.

Spearman correlation (full mapping score only) As done in Roads and Love (2020), one option in the case of the full mapping score, is to take the Spearman correlation between the upper trangular portions of the aligned pairwise distance matrices. Using the Spearman correlation could be yielding strong performance in part by minimising the impact of distributional differences in pairwise inter-concept distances (visualised in Figure 4.3).

Note that, in the case of Spearman correlation objective function, similarity relationship transformations do not make a difference to the alignment score. This is because all the transformation functions being tested are monotonic and therefore the ranks of distances do not change with transformation, and

so nor does the Spearman correlation. Hence, the test slate shown in Table 4.1 only includes one combination of Distance transformation × Objective $f(\cdot)$.

Pearson correlation (full mapping score only) An alternative to the Spearman correlation of pairwise distances would be to take the Pearson correlation between the upper-diagonal portions of two systems' similarity matrices, where the order of concepts in the matrices is dictated by the mapping between systems. This allows for us to explore the impact of similarity relationship transformations on the utility of the alignment score, incorporating our findings that close-range relationships may be more useful for alignment than long-range relationships.

Euclidean distance (pairwise score only) When using the pairwise score (see Figure 4.5), one option for computing the cost of matching a pair of items across systems is to take the Euclidean distance between the fingerprint vectors of the two items being paired. This is the default distance metric for this score type.

Wasserstein distance (pairwise score only) An alternative distance measure between the fingerprint vectors for a given pair of concepts is the Wasserstein distance. In its original formulation, the Wasserstein distance (or Earth Mover's distance) is a distance between two probability distributions, quantifying the minimum cost of turning one probability distribution into another. For the current problem, this has the potential to address weaknesses in the Euclidean measure. For example, the Wasserstein distance may not be as sensitive to small perturbations between systems as the Euclidean distance measure, as it incorporates an appreciation of column proximity, as opposed to treating each column of the fingerprint matrix as an orthogonal dimension.

Now that the objective functions that we are aiming to optimise have been established, we move on to the testing procedure, where we hope to establish which modifications are most effective for each function type.

Comparing candidate objective functions

To test how well an objective function corresponded to success in the unsupervised alignment task, we used the conditional sampling procedure established in Roads and Love (2020). This tests the relationship between the output of the objective function (the alignment score) and the mapping accuracy.

Method: Sets of $N \in \{10, 50\}$ concepts were sampled from the concepts in the intersection of the word and image embeddings being used. For each possible accuracy level, where $0 \leq n_{acc} \leq N - 2$, we sample $min(n_{perm}, 10,000)$ cross-system mappings from the set of n_{perm} possible mapping permutations where n_{acc} concepts are mapped correctly.

To explore the performance of these new settings, we look at two metrics: the logarithm of estimated proportion of misleading mappings when using the score ($log(p_{mm})$), and the Spearman correlation between the score and accuracy of a mapping (ρ). The objective of the test in this section is to find a scoring system which maximises ρ and minimises $log(p_{mm})$, as a stronger correspondence between the scoring system and accuracy lends itself to maximal accuracy of an algorithm which optimises the score in question.

Figure 4.2.2 plots score vs mapping accuracy for the scoring settings used in Roads and Love (2020). This setting applies no dimensionality reduction, no similarity relationship transformation, and uses a Spearman correlation objective function. This setting functions as our baseline for the full mapping score. In Table 4.1, this setting's ID is 2_S.

The red area in Figure 4.2.2 highlights sampled mappings which were misleading according to the score in question. That is, the score for these mappings was higher than the score for the correct mapping, despite their accuracies of course being lower. The metric $log(p_{mm})$ is the natural logarithm of the expected proportion of all possible mappings which would be misleading, based on the proportions of misleading mappings observed in the sampling process, and therefore captures the size of this red area.

Results: What follows are two tables which show the best performance of

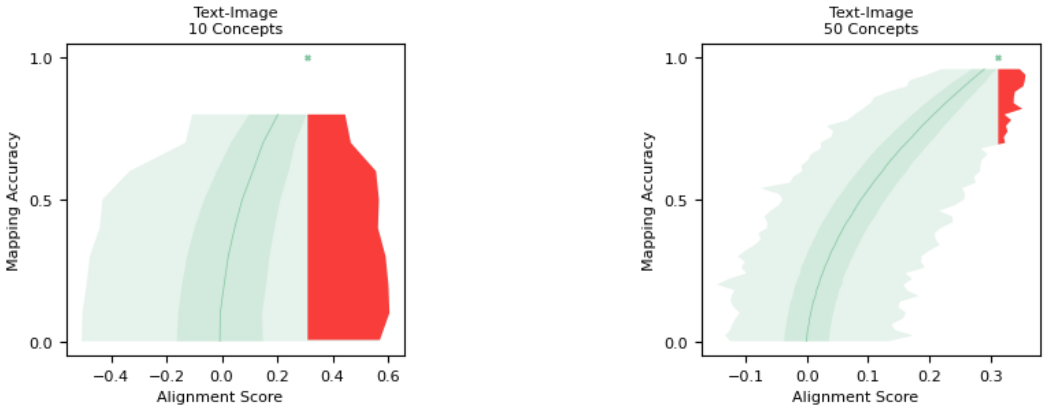


Figure 4.6: Plots of the relationship between alignment score and mapping accuracy for the settings used in the original Roads and Love (2020) paper. The left panel shows the relationship for 10-concept samples; the right panel shows the relationship for 50-concept samples In both plots, dark shading represents one standard deviation from the mean (represented by the dark line). Lighter shading shows the minimum-maximum envelope. Red regions represent misleading mappings. In this section, we attempt to identify scoring methods which can improve upon these baselines, in terms of the relationship strength and the number of misleading mappings..

any hyperparameter set for each setting combination. Table 4.1 summarises the results for scoring settings which are used for full mappings, and Table 4.2 summarises results for individual pairing scoring settings. Within each table the best performance on each metric is highlighted in bold font. Note that, while CCA is included in both tables as a reference for near-ceiling performance, it is not included in the selection of the best performance, as it is a supervised method.

The aim here is to identify a combination of settings for each algorithm type which outperforms the relevant baseline, by achieving a stronger relationship between accuracy and score – higher ρ – or a smaller number of misleading mappings – lower $log(p_{mm})$. For the full mapping score, shown in Table 4.1, the baseline score settings have ID 2_S. For the pairwise mapping score, shown in Table 4.2, the baseline score settings have ID 2_E. On the whole, we find that a UMAP dimensionality reduction and an exponential transformation of inter-concept distances are best able to improve the quality of the score. This demonstrates that the representation of similarity may be an important component of successful alignment algorithms, and that prioritising local relationships yields success.

The best combination of tested hyperparameters for each setup on each metric was selected on each of the 4 metrics in question (ρ and $log(p_{mm})$ for

ID	Dimensionality Reduction				Distance transformation			Objective f(·) Correlation		Performance N=10		N=50	
	None (0)	CCA[1] (1)	PCA (1)	UMAP (2)	None (0)	Exp (2)	Exp Rank (1)	r_p (0)	r_s (0)	ρ	$log(p_{mm})$	ρ	$log(p_{mm})$
1_P		✓			✓			✓		0.542	$-\infty$	0.964	$-\infty$
2_S	✓				✓				✓	0.205	-3.60	0.895	-102.32
2_P	✓				✓			✓		0.123	-3.15	0.920	-127.98
3_P	✓					✓		✓		0.427	$-\infty$	0.933	**-135.39**
4_P	✓						✓	✓		0.333	-7.03	0.914	-116.79
5_S			✓		✓				✓	0.381	-10.61	0.888	-102.29
5_P			✓		✓			✓		0.247	-7.404	0.891	-113.18
6_P			✓			✓		✓		0.441	-15.10	0.935	-134.75
7_P			✓				✓	✓		0.361	-11.51	0.928	-124.24
8_S				✓	✓				✓	0.423	-10.95	0.929	-120.50
8_P				✓	✓			✓		0.417	-11.96	0.922	-116.09
9_P				✓		✓		✓		0.501	$-\infty$	**0.937**	-120.51
10_P				✓			✓	✓		**0.506**	-13.56	0.929	-123.11

Table 4.1: Performance of all tested combinations of settings for the full mapping score. Cell values represent performance for the best combination of hyperparameters tested in a given row. r_p is Pearson correlation, r_s is Spearman correlation. Note that not only one distance transformation is tested when the objective function is Spearman correlation, as all transformations are monotonic and therefore the Spearman correlation renders the transformation meaningless in the eyes of the objective function.

ID	Dimensionality Reduction				Distance transformation			Objective f(·) Similarity		Performance N=10		N=50	
	None (0)	CCA[2] (1)	PCA (1)	UMAP (2)	None (0)	Exp (2)	Exp Rank (1)	d_E (0)	d_W (0)	ρ	$log(p_{mm})$	ρ	$log(p_{mm})$
1_E		✓			✓			✓		0.716	-13.06	0.949	-130.31
2_E	✓				✓			✓		-0.103	-0.392	0.607	-4.43
2_W	✓				✓				✓	-0.061	-0.47	-0.300	-0.19
3_E	✓					✓		✓		0.538	-7.36	0.740	-7.69
3_W	✓					✓			✓	0.453	-6.01	0.753	-8.56
4_E	✓						✓	✓		0.316	-2.87	0.577	-3.89
4_W	✓						✓		✓	0.475	-5.61	0.691	-6.02
5_E			✓		✓			✓		0.122	-1.24	0.501	-3.08
5_W			✓		✓				✓	0.365	-3.76	0.034	-0.78
6_E			✓			✓		✓		0.431	-5.02	0.653	-5.26
6_W			✓			✓			✓	0.419	-4.31	0.510	-3.16
7_E			✓				✓	✓		0.354	-4.16	0.602	-4.36
7_W			✓				✓		✓	0.436	-5.24	0.221	-1.33
8_E				✓	✓			✓		0.608	-12.51	0.889	-58.68
8_W				✓	✓				✓	0.456	-9.65	0.788	-10.34
9_E				✓		✓		✓		**0.698**	**-14.00**	**0.938**	**-95.24**
9_W				✓		✓			✓	0.637	-12.22	0.892	-63.61
10_E				✓			✓	✓		0.689	-12.79	0.911	-71.47
10_W				✓			✓		✓	0.594	-11.02	0.889	-49.55

Table 4.2: Performance of all tested combinations of settings for the individual mapping score. Cell values represent performance for the best combination of hyperparameters tested in a given row. d_E is Euclidean distance, d_W is Wasserstein distance

each of N=10 and N=50). The Spearman correlation between ρ and $log(p_{mm})$ was 0.975 ($p < 0.001$).

For reference, the score v. accuracy plots for CCA - the supervised comparison - is included below in Figure 4.7 for the full mapping score, and Figure 4.9 for the individual mapping score. As is visible from both the table and the plots below, CCA generated no misleading mappings for either set size in the full case, and had the highest values of ρ for both set sizes as well. In the individual case, CCA had the best performance on ρ and $log(p_{mm})$ for N=50, but was outperformed by another setting on $log(p_{mm})$ for N=10.

Plots for the best performing parameters for the full mapping score are shown in Figure 4.8. Plots for the best performing parameters on the individual mapping score are shown in Figure 4.10. In both cases, we observe a strong relationship between score and accuracy, and find combinations of settings which outperform the baselines in Roads and Love (2020). For the full mapping score, the best settings use dimensionality reduction and/or transformation of similarity relationships, depending on the measure of success. For 10 concepts, the best performing settings are 10_P and 9_P respectively; for 50 concepts, the best performing settings are 9_P and 3_P.

For the individual mappings, we had no starting point from prior work, but similarly we observe that the best combination of settings includes a dimensionality reduction and distance transformation. For both set sizes and on both measures, the best setting is 9_E.

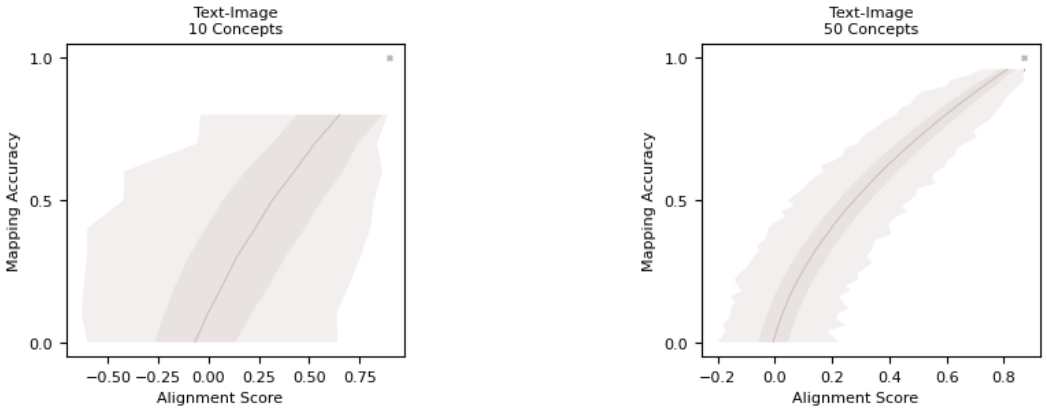


Figure 4.7: Score performance with CCA dimensionality reduction (full mapping). This serves as a proof of concept for the possible impact of dimensionality reduction methods/space manipulations on the ability for an alignment score to help perform alignment, but is not performed in an unsupervised fashion.

4.2.3 Alignment algorithms

Having identified modifications to the alignment score which improve the possibility that it will guide us to accurate mappings, I now proceed to explore candidate alignment algorithms for the unsupervised mapping problem.

Chapter 3 demonstrated success in using a CycleGAN inspired (Zhu et al., 2017) model to fit human behaviour in a paired-associate learning task. But could a different, less human-like model do a better job of alignment than

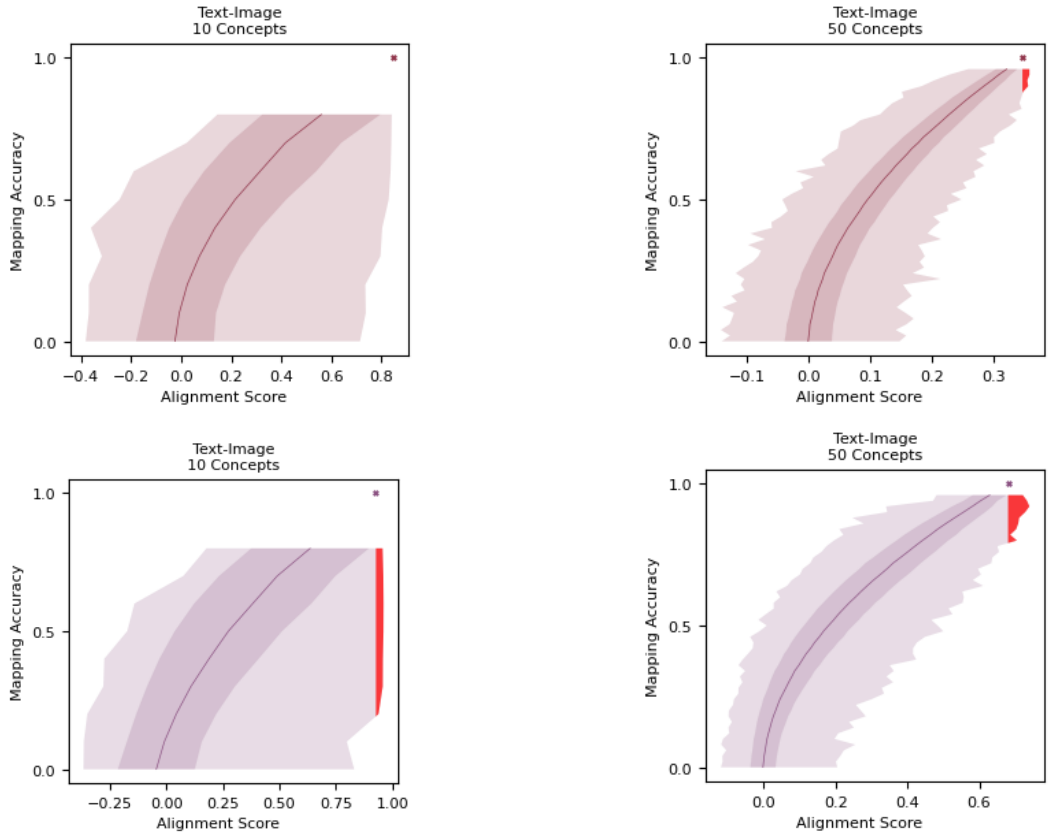


Figure 4.8: Plots of alignment score (full mapping) performance for the score settings which optimised the number of misleading mappings (top row) and the correlation between score and accuracy (bottom row). For 10 concepts, the best performing settings are 10_P and 9_P respectively; for 50 concepts, the best performing settings are 9_P and 3_P.

this cycle model? Or does this model indeed yield superior alignment results compared to other candidate approaches? To begin answering this question, this chapter evaluates the following classes of alignment algorithm, which were selected to provide a good spread across suitable approaches identified from prior work:

- **The Kuhn-Munkres algorithm** (or, Hungarian algorithm): an optimal transport method
- **The Cycle model**: an extension of the CycleGAN-inspired aligner used in the behavioural experiment (Zhu et al., 2017)
- **Monte Carlo Tree Search (MCTS)** (Browne et al., 2012; Pinheiro et al., 2016),

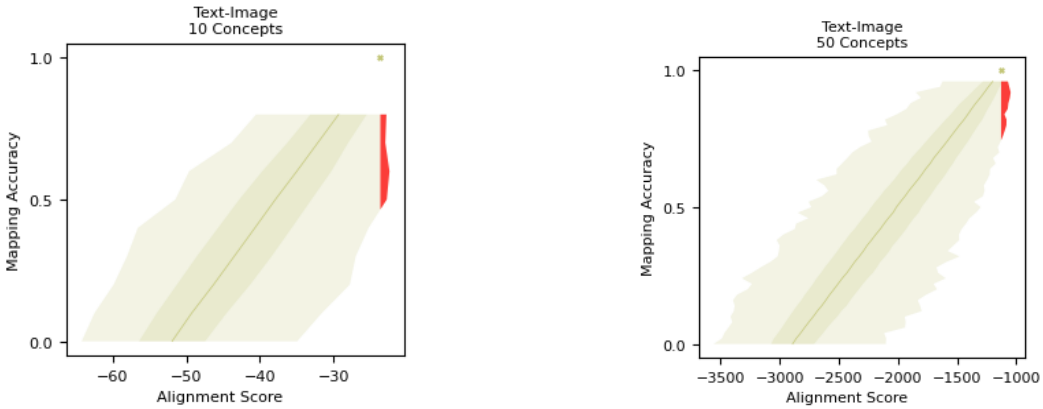


Figure 4.9: Score performance with CCA dimensionality reduction (individual mappings). This serves as a proof of concept for the possible impact of dimensionality reduction methods/space manipulations on the ability for an alignment score to help perform alignment, but is not performed in an unsupervised fashion.

Kuhn-Munkres algorithm

The Kuhn-Munkres algorithm is an algorithm for the one-to-one assignment of items in one system to items in another system, such that the final assignment minimises the overall cost of assignments. In its original formulation, system X consisted of 'workers' and system Y of 'jobs', where each worker had a cost associated with each job. The assignment problem which the Hungarian algorithm solves is formalised as follows: for two systems X and Y of equal size and a cost matrix $C : X \times Y \to \mathbb{R}$, find the bijection $f : X \to Y$ which minimises the cost function $\sum_{x \in X} C(x, f(x))$ (Kuhn, 1955).

In our case, system X is the embeddings in one modality and system Y embeddings in the other. The cost matrix C is calculated as the pairwise distance matrix between the fingerprint matrices of items in systems X and Y, where fingerprint matrices are constructed by sorting the in-system similarity relationships row-wise. This is visualised in Figure 4.5. The cost matrix is calculated as $C_{alignment} = 1 - \frac{1}{2}(corr + 1)$, such that all cost values $\in [0, 1]$. Details of the Hungarian algorithm are included in Appendix C1.2

Note that in this unsupervised form, using fingerprint distances as a cost function (see Figure 4.5), the Hungarian algorithm does not actively seek to optimise for alignment, as mappings are made on an individual basis using unaligned patterns of relationships. It is therefore agnostic to global structures and to the consistency of mapping decisions across the set of points. But in the

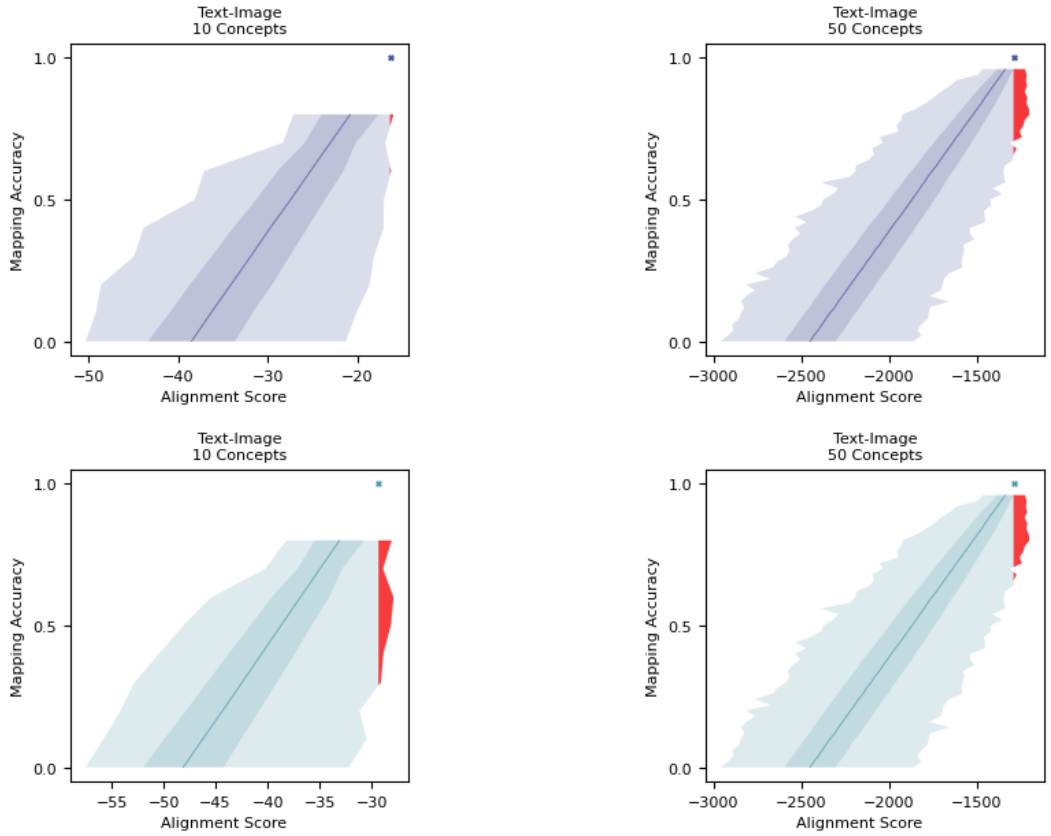


Figure 4.10: Plots of alignment score (individual mapping) performance for the score settings which optimised the number of misleading mappings (top row) and the correlation between score and accuracy (bottom row). For both set sizes and on both measures, the best setting is 9_E.

trivial case where all relationships were identical across systems, this algorithm is well placed to solve the task.

In turn, with some relationships known, the cost matrix can be adapted to use alignment as a means of performing cross-system mappings. For example, with a set of concepts Y_s whose mappings are known, the cost matrix C can be re-formulated such that the algorithm maximises the correlation between similarity relationships to known items in systems X and Y for the selected matched pairs. This is discussed in more detail in section 4.2.4.

Cycle

The structure of the cycle model used here was very similar to the Aligner component of the Regression + Aligner model used in Section 3.3. The model was comprised of two multi-layer perceptrons (MLPs): $F(.)$ and $G(.)$. One performed mapping $F : X \rightarrow Y$ from space X into space Y, and the other

mapped in the opposite direction $G : Y \to X$. The MLPs used in this chapter had 2 hidden layers of size 100. This model used only the unsupervised components of the loss term: cycle consistency loss and distribution loss. λ_{cyc} was fixed at 1. λ_{dist} and the standard deviations of the multidimensional Gaussian kernels in the distribution loss, σ, were hyperparameters. Training was conducted over 1000 iterations. An Adam optimizer (Kingma and Ba, 2014) was used with a learning rate of 0.001.

Supervision can be incorporated by re-introducing the regression component of the loss. This is discussed further in 4.2.4.

MCTS

Monte Carlo Tree Search is a search method which combines tree search with random sampling, enabling the search of vast search spaces with some notable successes. The process of MCTS is summarised in Figure 4.11, (Browne et al., 2012). The algorithm builds a tree by simulating series of actions each time it reached an unvisited node, which are used to estimate the values of tree nodes. Before the simulation begins, nodes are selected at each level of the tree based on the selection policy. Once the simulation has begun, nodes are relected based on the rollout policy. At the end of each simulation, we calculate the value of the mapping using the alignment score described above. This score is backpropagated to the node from which simulation began.

To apply MCTS to the search for a mapping between spaces, we frame the problem as a single-player game (Schadd et al., 2008), where each node in the tree is defined as the mapping between an unmapped point in system X and one in system Y, given all previous points mapped across the systems.

We explore some variations of MCTS to best understand its scope for success in this problem. Variations occur in two places: the selection policy and the rollout policy. The variations we test in each policy are outlined below.

Selection policy

- **UCB1**: The UCB1 score which weighs value estimates derived from

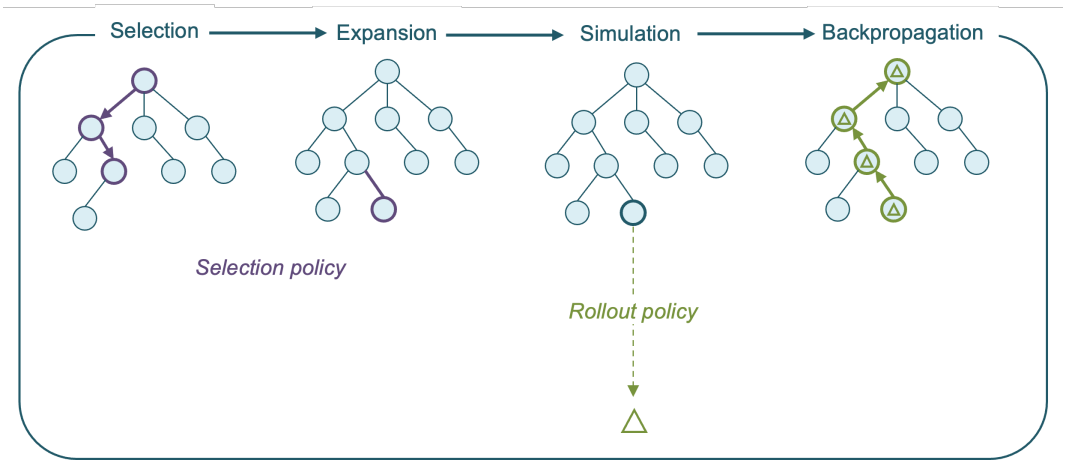


Figure 4.11: Overview of the MCTS algorithm, adapted from Browne et al. (2012)

simulations against the number of times each node has been visited. MCTS algorithms whose tree policies are based on the UCB1 score are commonly referred to as *Upper Confidence Bound for Trees* (UCT) algorithms (Browne et al., 2012). Details of the UCT algorithm are provided in Appendix C1.1. Schadd et al. (2008) successfully implements a modified version of the UCB1 score for single player games, given that single-player games do not have a win/loss/draw outcome. Here, for example, the outcome of our 'game' is an alignment score. We use this modified UCB1 score given in Equation C1.2 (Appendix C1.1) in our UCT algorithm.

- **Alignment heuristic**: For a large problem, like the mapping between two concept spaces, domain knowledge may improve the efficiency of tree search. In our case, we know that in order for the alignment score to be maximised for a complete mapping, alignment scores of mappings made along the way will for the most part be maximised (or close to maximised). We also know that alignment scores are more likely to be reliable where they are calculated with respect to a larger number of concepts. Therefore, we implement a version of the MCTS algorithm which uses an alignment heuristic for node selection.

  This heuristic operates by selecting nodes on how the candidate node impacts the alignment score for the current mapping. Taking the parent

node as a fixed mapping to the original system, for items 0,...,i-1, what is the alignment score when each candidate node is chosen as a match for node i? We take the softmax of the candidate nodes' alignment scores to generate a probability distribution across candidate matches for node i. The temperature parameter T in the softmax function decreases (i.e, makes the distribution less uniform) as i increases, owing to the knowledge that alignment scores for more known items are more reliable.

Values of visited states are tracked, and an average is updated as in UCB1, to account for the possibility that some missteps are made by greedily maximising alignment score. This should identify any missteps which yield lower alignment scores in the final mapping, for example by filling a spot early on when it is better suited to a future spot, thus possibly avoiding local minima.

- **Exhaustive start:** There is a strong possibility that seeding a system with a small number of correctly mapped concepts would aid performance. Alignment could then proceed based on relationships with known concepts. As such, we investigated a version of MCTS which upweighted the top of the tree - prioritising finding a small set of mapped concepts which led to superior mappings further down the tree, and then using these to anchor further decisions in the tree once they were locked in. The details of how this is implemented are found in Appendix C1.3

  After the pre-determined number of items have been searched exhaustively, we revert to the alignment heuristic.

Rollout policy

- **Random:** The random rollout policy randomly selects states in simulation phase. While this method may work for small problems, it may become too difficult to find the correct answer as the problem size increases.

- **Alignment-constrained** According to this policy, the probability of

choice is determined by the alignment score with respect to currently mapped items. A softmax function is used to generate this probability distribution from alignment scores, and the temperature of this softmax function is decreased as the number of concepts increases. The rationale for this is that the alignment score becomes more reliable when a larger number of items have already been mapped. The temperature decreases linearly between a max of 1 and a min of 0.2 with each level of the tree.

Top-1 accuracy

To evaluate the alignment algorithms, the top-1 accuracy is used. This measures the percentage of mappings which are mapped to their correct counterpart across systems. MCTS and the Kuhn-Munkres algorithm both naturally output a one-to-one mapping between visual and linguistic entities. For the Cycle model, the one-to-one mapping is generated by using the Sinkhorn algorithm. Once the mapping fucntion $F : X \rightarrow Y$ has been trained, the Sinkhorn algorithm is applied to the pairwise distances between F(X) and Y in order to find the optimal pairwise correspondence.

4.2.4 Algorithm testing

Throughout this section, horizontal green lines on plots are used to visualise performance at chance in each testing condition.

We start by testing the algorithms on artificial, noise-free unsupervised mapping problems, from which the problems gradually become more realistic. Certain algorithms which were successful for perfect mapping conditions become very poor performers for naturalistic cross-modal mapping problems. We find that the most computationally expensive variant of MCTS is the most successful on the unsupervised mapping problem, but that in general it is very difficult for algorithms to achieve efficient success in this problem.

Having demonstrated the difficulty of unsupervised mapping, we add some supervision to the mapping problem to test algorithm performance in this case,

and find that the Cycle + Regression model and the Kuhn-Munkres algorithm are the most successful when some concept mappings are known.

Self-self mapping as proof of concept

To begin, and to explore the theoretical efficacy of the candidate algorithms for the task of alignment, we run the alignment algorithms on a test slate of self-to-self mappings. In the first instance, system X and system Y are both sets of 50-dimensional word embeddings sampled from the 418 concepts in the word-image embedding intersection. We test the algorithms on the sample sizes $N_s \in [10, 50]$, to explore whether the number of candidate items leads to changes in performance. While this problem may seem trivial, as it is a completely noise-free mapping, and therefore very unrealistic for something like cross-modal alignment, the journey from this trivial problem to the more complex solution may highlight individual weaknesses or strengths of different algorithm types. Thus we take a systematic approach from this trivial problem to the full cross-modal problem.

The results for the self-self mapping test are shown in Figure 4.12. The Kuhn-Munkres algorithm achieves 100% accuracy in this mapping for all problem sizes. The most computationally intensive MCTS variant - exhaustive start selection policy + alignment-constrained rollout - is the next best performer. These are the only two algorithms whose performance is significantly different to chance on both problem sizes when tested with post-hoc Bonferroni-corrected t-tests (see Appendix C2.1). The heuristic-constrained MCTS variant performs significantly better than chance for the 10 concept problem in the same test, but fails in the 50-concept problem.

The impact of noise

Next, some noise is introduced to the problem, which immediately impacts the performance of the Kuhn-Munkres algorithm. Different levels of noise (ϵ) are explored, to determine whether algorithms are differentially sensitive to perturbations in the matched embeddings. We add noise to the self-self mapping

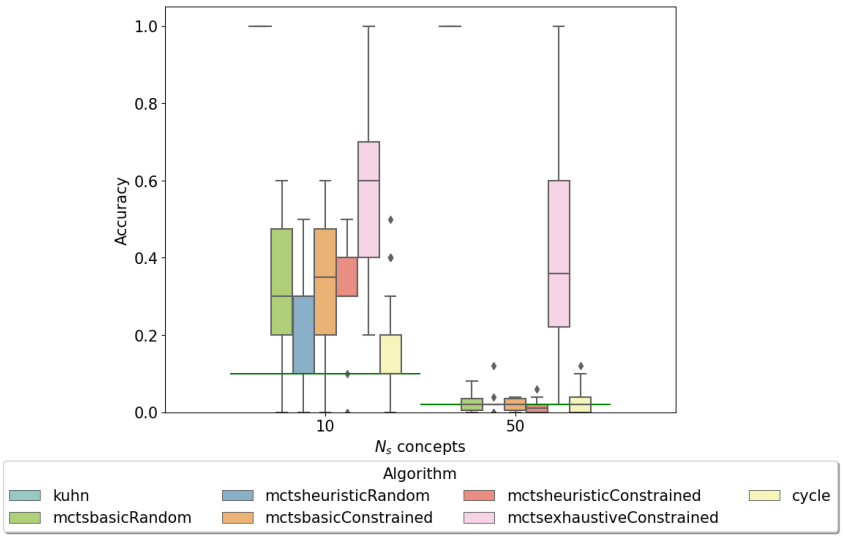


Figure 4.12: Results for a self-self mapping of 50-dimensional word embeddings. This tests the algorithms in a completely noise-free environment, which is not realistic, but demonstrates algorithm performance under perfect conditions. Note that all runs of the Kuhn-Munkres algorithm achieve perfect mappings in this artificial case. As such, their performance in this plot is represented by a horizontal line at accuracy level 1.0.

problem by sampling noise for each point in system Y from a multivariate Normal distribution $X \sim \mathcal{N}(\mu, \Sigma)$, where $\mu = 0_{1,D}$ and $\Sigma = \epsilon I_D$. D is the number of dimensions in the embedding to which we are adding noise. Here, $D = 50$.

The results for mappings of noisy embeddings are shown in Figure 4.13. For the lower noise conditions, the Kuhn-Munkres algorithm is generally the highest performance solution, followed by the exhaustive/constrained variant of MCTS. However, its performance is substantially affected by noise, becoming comparable to the performance of other methods as noise increases. In the highest noise condition, the exhaustive-constrained MCTS variant is the only algorithm which achieves performance significantly above chance in both problem sizes (see Appendix C2.2 for statistical tests).

Naturalistic, different problem sizes

Moving from the self-self mapping to a more complex problem, we deploy the algorithms on the problem of finding an optimal mapping from word- to image-systems. Here, we test performance on three different configurations of the objective function: *baseline*, which corresponds to the Spearman correlation on mapped similarity relationships (as implemented in Roads and Love (2020));

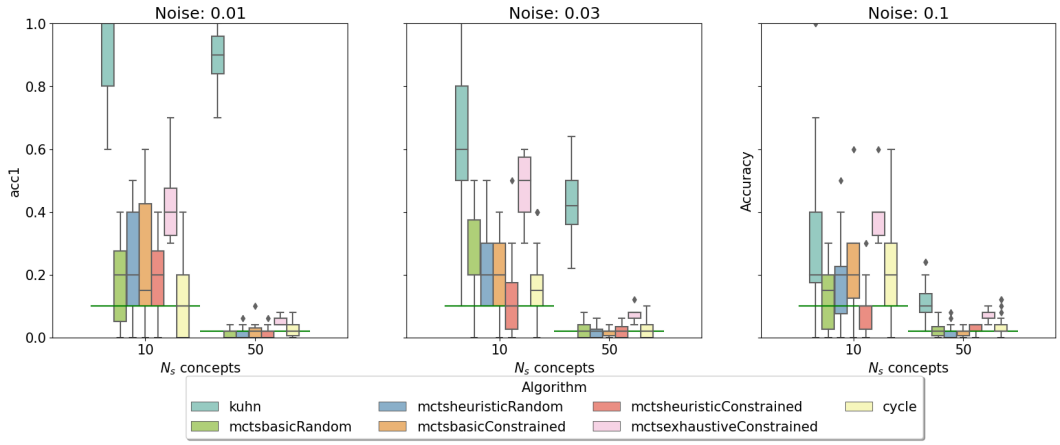


Figure 4.13: Results for a self-self mapping of 50-dimensional word embeddings, with artificial noise added to increase problem difficulty. This demonstrates the impact of noise on the performance of different algorithms on unsupervised alignment. Noise quickly impacts the performance of the Kuhn-Munkres algorithm, while other algorithm performance is more robust to the addition of noise.

bestProp, corresponding to the configuration which yielded the lowest proportion of misleading mappings in section 4.2.2, and *bestCorr* which yielded the highest correlation with mapping accuracy in section 4.2.2.

Figure 4.14 shows the performance of each system of interest on word-image mapping problems for 10- and 50- concept problems. Having moved to the cross-modal mapping problem, the performance of the Kuhn-Munkres algorithm becomes indistinguishable from chance (see Appendix C2.3 for statistical tests). So while this algorithm was successful for synthetic problems, it fails in the face of the noise introduced by the cross-modal mapping problem.

The computationally intensive exhaustive search variant of MCTS is the best performer on the real unsupervised task. This is perhaps not surprising, given that we know the alignment signal exists, and have shown in section 4.2.2 that maximising the optimised objective function should in theory yield success in mapping. The exhaustive MCTS variant is the algorithm on the test slate which uses the highest computational budget to search through mappings, and is thus able to deliver the best results in the unsupervised task. No other algorithm delivers performance significantly different to chance on any setting or problem size in the unsupervised cross-modal mapping problem.

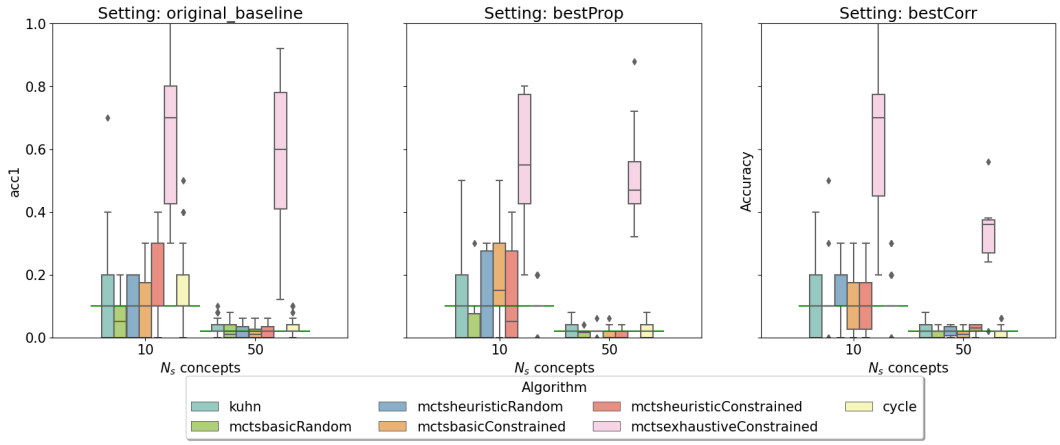


Figure 4.14: Results for a word-image mapping for 50-dimensional word embeddings and 10-dimensional image embeddings. First row is top 1 accuracy; second row is top 5 accuracy. Green dashed line shows chance performance

Introducing supervision

Given the challenges encountered in attempting the fully unsupervised problem, the following test sought explore how these algorithms could perform when some cross-modal mappings were known. First, the way in which supervision is incorporated into each algorithm is explained. Then, the results are presented.

Supervision in Kuhn-Munkres algorithm To incorporate supervision into the Kuhn-Munkres algorithm, the cost matrix is updated, such that minimising the cost corresponds to finding a mapping which matches items with similar relationships to the known concepts in each modality. To do this, the entries to the cost matrix $C(i, j)$ are the inverse of the alignment score obtained when x_i is mapped to y_j, given the known items $y \in Y_s$ in system Y and $x \in X_s$. This is visualised in Figure 4.15 (such that minimising the cost function corresponds to maximising the alignment score).

Supervision in Monte Carlo Tree Search Incorporating supervision into MCTS, the alignment score calculated at the end of a simulation is calculated with respect to the known concepts. This is then backpropagated to update the values of the nodes that were visited before simulation began. Furthermore, for MCTS variations using the alignment heuristic, the probability distributions across concepts as the algorithm moves through the tree are now calculated using the alignment scores relative to the known concepts.

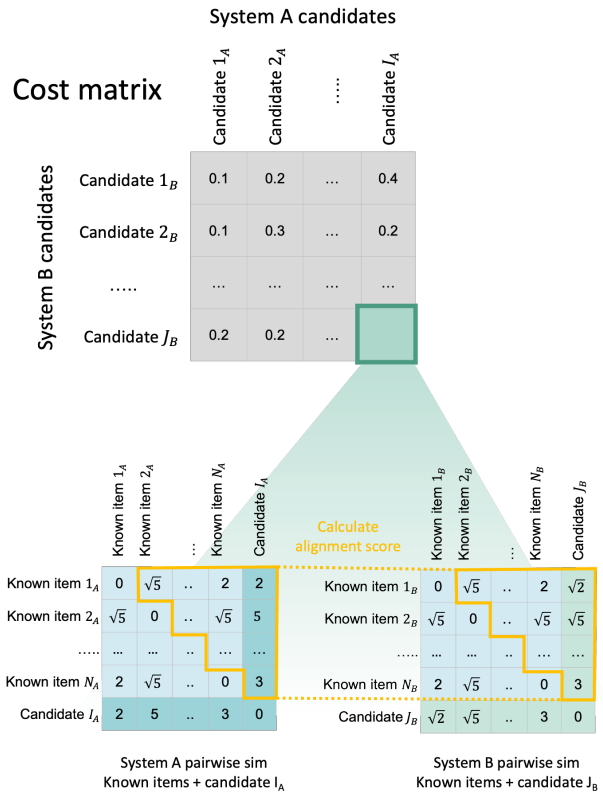


Figure 4.15: Schematic of how supervision signals are incorporated into the Kuhn-Munkres algorithm.

As the exhaustive search MCTS variation was intended to be a proxy for supervision (by finding successful early mappings to anchor the rest of the search), this variation is now dropped in the presence of supervised items.

Supervision in Cycle In the Cycle model, supervision is incorporated through an additional term added to loss function. This additional loss term, λ_{sup}, is the sum of the distances between $F(X_i)$ and the corresponding Y_i for $i \in S$, where S is the set of seen concepts (that is, items for which the mapping is known). In other words, on top of optimising the unsupervised cycle loss, the model now also aims to accommodate the known relationships across modalities.

For a fair comparison, a Regression model is also added to the model slate. This is a model with the same underlying structure are the Cycle model, in that it is comprised of mapping functions F(.) and G(.) which map $F : X \rightarrow Y$ and $Y : Y \rightarrow X$ respectively. However, the Regression model does not include the unsupervised alignment term in its loss function, and instead only minimises the loss λ_{sup} - the distance between mapped items and their known

counterparts.

4.2.5 Supervision experiments

The results for tests of the effect of supervision are shown in Figure 4.16. Supervised concepts are randomly selected from the set of concepts in the embedding intersection. The Kuhn-Munkres algorithm begins to perform better than chance with 10 supervised examples. With 50-supervised, both the Kuhn-Munkres algorithm and the Cycle + Regression models outperform chance on both problem sizes (see Appendix C2.4 for statistical tests). This agrees with the intuition and findings from Roads and Love (2020), that the more mappings are known the easier learning by alignment can become. Interestingly, the Regression only model does not perform as well as the Cycle + Regression model, and does not in any test configuration outperform chance.

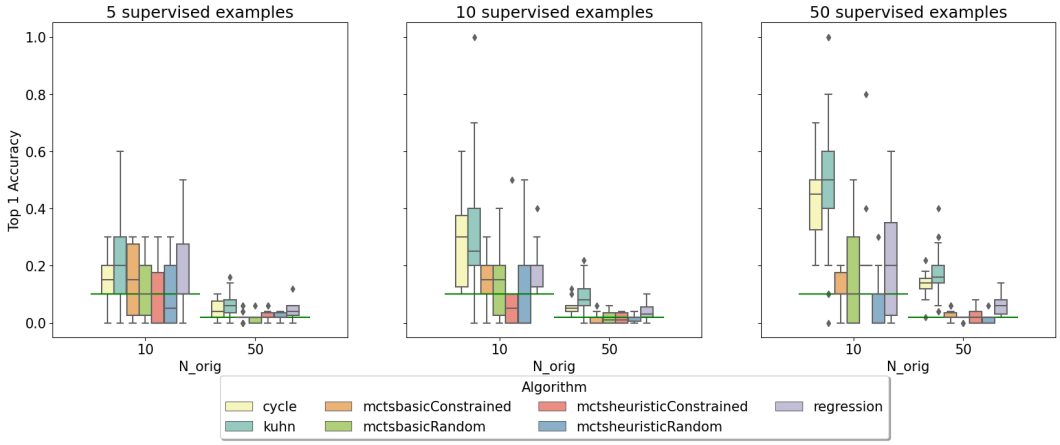


Figure 4.16: Plot to show algorithm performance with various levels of supervision (known concepts).

4.3 Evaluating alignment as prior

An idea introduced in Chapter 2 was that systems alignment could function as a prior for learning object-word mappings in the real world. In machine learning contexts, learning object-word mappings corresponds to the task of image classification. The following section examines whether systems alignment could serve as a valuable prior in object classification tasks.

Based on the findings of the previous section, we are particularly interested in how knowledge of some classes can be leveraged to generate priors across novel classes (Socher et al., 2013; Frome et al., 2013; Akata et al., 2015). In line with the terminology established in Socher et al. (2013), we refer to the classes on which the prior is trained as seen classes Y_s, and the novel classes on which classification is tested as unseen classes Y_u.

We are interested in exploring whether alignment processes based on known classes better enable a learner to correctly generalise its knowledge to new categories with little to no supervision. This is akin to a real-world learning scenario where some concepts have been learned based on multiple supervised instances, and other classes remain unknown.

Learning scenarios where the availability of supervised examples for a classification is low are often known as *few-shot learning* scenarios. This could be considered a machine-learning parallel to the naturalistic learning environments of a human learner, where few explicitly supervised examples of object labelling are encountered (Tamis-LeMonda et al., 2019; Clerkin et al., 2017). We use a few-shot learning test to see how well learners can learn new classes from a small number of supervised instances. We also test zero-shot learning, to see how well the learner can generalise to new classes with no supervised examples at all. Details on both few- and zero-shot learning are provided below.

A key question in this exploration is whether the inclusion of unsupervised alignment information benefits performance of a cross-system mapping model. This is of interest because the model that was identified in Chapter 3 as a the best fit for human learning included an unsupervised component alongside the supervised learning component. Additionally, the Cycle model was a top performer when a sufficient number of supervised classes were provided to the system in the previous section.

In the coming section, the model which contains an unsupervised alignment component engages this component while learning its seen classes Y_s. The

model is provided with some supervised examples from the known classes, for which an image is provided alongside its label, but also with unsupervised examples from these classes. We test whether this model, which attempts to learn from these unsupervised examples, is able to outperform a comparison model which relies on the supervised examples alone.

This reflects real-world learning scenarios: while some synchronous cross-modal examples will exist for a class of objects (where a member of the class is seen and the accompanying label is provided at the same time), there will be many examples within a class which are seen, but without the appropriate label being provided. This exploration is also highly relevant to machine learning contexts, where labelled data is hard to come by. If unsupervised learning processes prove valuable for establishing a cross-modal mapping, this could allow machine learning systems to make use of unlabelled data to facilitate learning.

Few-shot learning

Few-shot learning (FSL) is the task of learning to perform a task - here, object classification - after training on only a few examples (Wang et al., 2020). Humans are capable of performing few-shot learning of categories, and can generalise to new objects successfully (Potter, 1976; Thorpe et al., 1996).

One approach to few-shot learning is to perform unsupervised pre-training on unlabelled data, followed by fine-tuning using the small number of labels available within the FSL environment (Chen et al., 2020c). The unsupervised pre-training step makes the most of unlabelled data by leveraging it in a task-agnostic fashion. This approach has been highly fruitful in natural language processing contexts (Devlin et al., 2018), and has more recently been applied to the task of image classification within the computer vision domain (Chen et al., 2020b,c).

If n images per class are used for classifier training, this is referred to as n-shot learning (e.g, if 2 images per class are used for the training set, this is

2-shot learning).

Zero-shot learning

By extension, zero-shot learning is performing a task with no training examples present (Pourpanah et al., 2022). As a result, zero-shot learning tasks require some form of 'background' information in order to be successful, such as semantic information or other context. Prior studies have shown that by incorporating semantic representations for class labels, zero-shot mapping of images to previously unseen class labels is possible (Socher et al., 2013; Frome et al., 2013; Akata et al., 2015).

SimCLR embeddings

For this assessment, we use visual embeddings extracted from SimCLRv2 (Chen et al., 2020c). SimCLR, the 'simple framework for contrastive learning of visual representations', is a self-supervised embedding algorithm, which generates representations based on objects' visual features (Chen et al., 2020b). Representations are trained by encouraging the embedding model to maximise agreement between representations for transformations of a single image, while simultaneously maintaining distance between an image and other images.

While the SimCLR embeddings themselves are trained in an unsupervised fashion, and are therefore task-agnostic, the performance of SimCLR embeddings on downstream tasks such as image classification is highly impressive. SimCLR embeddings outperform both fully-supervised classifiers and state-of-the-art semi-supervised classifiers on image classification tasks with small numbers of training examples (i.e, in few-shot learning environments) (Chen et al., 2020c).

A key difference between SimCLR embeddings and the co-occurrence based visual embeddings used previously, is that SimCLR embeddings are inferred for raw image inputs. When it comes to evaluating the potential for alignment to benefit machine learning systems, this provides a valuable testing ground,

as there is no pre-requisite for category labels to be inputted into the visual system before the training for alignment can take place.

One consequence of this is that the alignment problem becomes a many-to-one mapping problem. In the visual co-occurrence based embeddings used so far in this thesis, the mapping problem was one-to-one, as visual object categories were being aligned with linguistic category labels. But using SimCLR, it is individual visual images which are mapped onto linguistic labels, and there are many visual images associated with each potential label.

Adding an alignment prior

Can a machine learning system leverage existing knowledge of cross-modal mappings to improve performance on novel image classes? The scenario explored here is as follows: if a machine learning system has knowledge of how some items relate to their class labels, but no such knowledge for any examples in other classes, can the novel classes be learned more effectively by extracting a prior distribution across classes based on alignment? If so, are some alignment mechanisms more successful at generating a successful prior than others?

Based on the findings presented in this thesis, it is plausible that alignment information would be able to improve performance on image classification when there is little supervisory data available. In Chapter 2, it was shown that novel image-word mappings could be inferred based on known concepts using alignment principles. In Chapter 3, in a completely supervised task, learning efficiency and final performance in humans were both improved by the mere presence of alignable systems. In the first section of the current chapter, semi-supervised alignment algorithms were found to be beneficial for highly computationally complex image-word labelling tasks. Could the findings of this thesis' previous chapters and the algorithmic exploration above be combined to contribute to successful learning in a few-shot learning environment?

The process for the incorporation of an alignment prior into the classification pipeline is visualised in Figure 4.17. The pipeline proceeds as follows: the image classification task is restricted to Y_u 'unseen' image classes. Image representations resulting from unsupervised pre-training are obtained from SimCLR, for n images per novel image classes (for 1-shot learning. $n = 1$; for 2-shot learning, $n = 2$ etc). Using an alignment-based method, trained on Y_s 'seen' image classes, a prior distribution across the Y_u unseen image classes is generated for each image $x_i \in X_y$ for $y \in Y_u$. Note that the sets of known image classes and novel image classes are mutually exclusive. Then, the linear classifier is trained on the small number of supervised examples in the training set, and the prior for each image is inputted into the linear classifier alongside its feature vector.

We compare 4 prior conditions: no prior information (Uniform prior); a Correlation-based prior; a Regression prior and a Cycle + Regression prior. The Correlation-based prior is conceptually similar to the underpinnings of the forced-choice task in Chapter 2, and to the pairwise cost function used in the Kuhn-Munkres algorithm in the previous section. The Regression and Cycle + Regression priors involve training mapping functions to map between visual and linguistic space using the known item classes, and then applying these trained mapping functions to the novel classes. These two prior types are derived in the same way, with the only exception being the loss term used to train the mapping models. More details on each model type are provided in section 4.3.3.

4.3.1 Materials

Visual representations using SimCLR

The images used in this study are sourced from the ImageNet ILSVRC2012 dataset. This dataset contains labelled images from 1000 object categories. For our purposes, it was necessary that the class names corresponded to a word with a GloVe embedding in the set described below. Some minor pre-processing

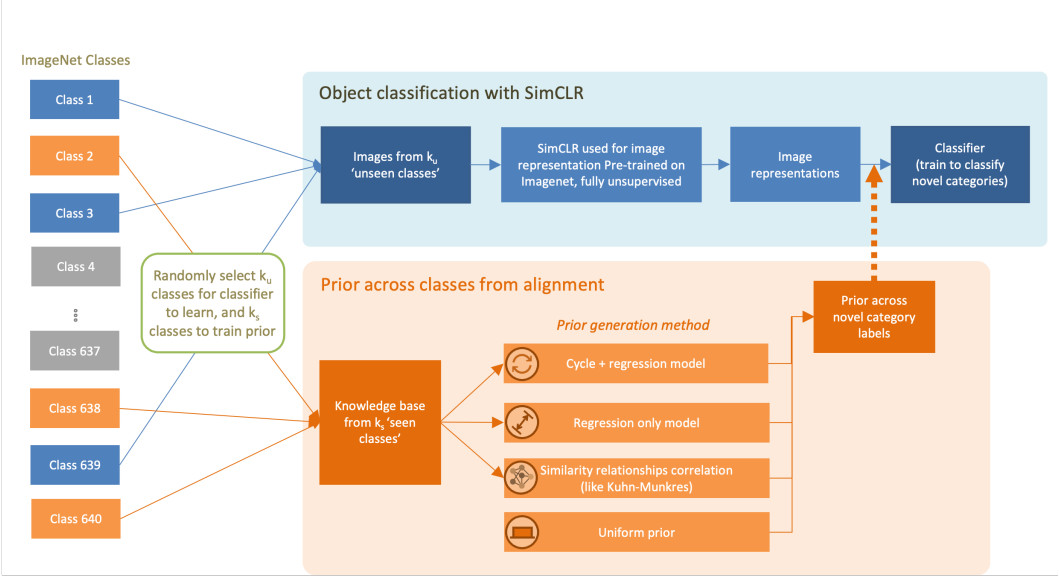


Figure 4.17: Schematic demonstrating how the prior interacts with the classifier in the classification task.

of class names done to align class names with GloVe embeddings (e.g., 'cellular telephone' ILSVRC2012 class name amended to 'cellphone', present in the set of GloVe embeddings). After pre-processing, the total number of classes whose names aligned with available GloVe embeddings was 640. This is the set of classes Y from which we select our seen and unseen categories, Y_s and Y_u respectively.

To obtain image representations, images are passed through the pretrained SimCLR model with a ResNet50 backbone (1x width, see Chen et al. (2020c)). From this, we obtained 2042-dimensional image representations.

In total, this resulted in a dataset of 822,982 image representations, with dimensionality 2042 image representations across 640 classes. The mean number of images per class was 1286, with the minimum number of images in any class being 754.

GloVe embeddings

The pre-trained word embeddings used here were 50-dimensional GloVe text embeddings (Pennington et al., 2014). These embeddings are trained on 6 billion tokens from the Wikipedia2014 + GigaWord5 text corpus. The resultant vocabulary size is 400,000 tokens.

4.3.2 Are SimCLR embeddings and GloVe embeddings alignable systems?

If alignment is going to be useful as a source of prior information for this task, a necessary pre-requisite is that SimCLR embeddings and category labels are alignable systems. In the prior sections of this thesis, visual embeddings based on the co-occurrences of objects in visual space have been used to probe the role of alignment in learning cross-modal mappings. The alignability of these systems was confirmed (Roads and Love, 2020), using a conditional sampling procedure which tested the alignment of cross-system mappings with different mapping accuracies. As described above, this section uses embeddings inferred using contrastive learning, based on the visual features of images, and are shifting to the use of SimCLR embeddings. Thus, it is worth testing the alignability of SimCLR embeddings and the GloVe embeddings for the associated class labels.

It is reasonable to believe that representations based on visual features and those based on language use might be alignable. As has been discussed in previous sections of this contribution, things which are spoken about in similar contexts are likely to be similar in form (for example, pens and pencils look alike because both are used for writing, and they are spoken about similarly for the same reasons). There is supporting evidence for this relationship in Johns and Jones (2012), where it was demonstrated that if the perceptual features of some object names are unknown, but the similarity relationships for the object names in the linguistic space are known, perceptual features can be inferred to an impressive degree of success using a simple associative mechanism.

To confirm the presence of alignable similarity relationships, I conducted a systems alignment analysis of these systems, using the conditional sampling procedure outlined in section 4.2.2.

A key amendment to the conditional sampling procedure arises from the many-to-one nature of this new problem. Specifically, SimCLR embedding positions were averaged within categories before alignability was tested using

conditional sampling. For each image class, the embedding positions were averaged across categorised items.

The results of this conditional sampling analysis are provided in Figure 4.18, confirming that the SimCLR embeddings and GloVe embeddings are indeed alignable.

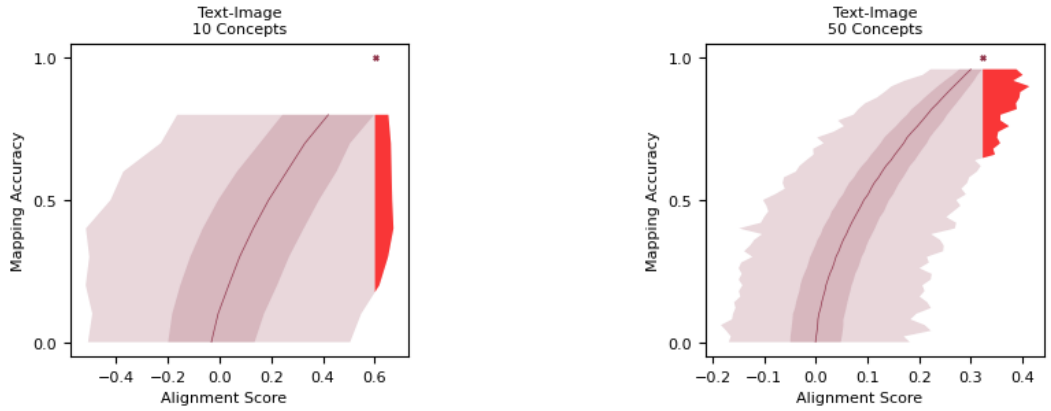


Figure 4.18: Conditional sampling analysis results for SimCLR embeddings and GloVe embeddings. SimCLR embeddings are averaged within classes to give a mean class position in image embedding space, and conditional sampling is performed on mappings between these mean positions in visual space and class labels in GloVe embedding space.

4.3.3 The impact of alignment priors

This investigation tested the impacts of different types of prior on few-shot classification. The following section describes the methods associated with (A) prior generation, and (B) classifier training, as outlined in Figure 4.17.

Prior conditions

We compare the performance of the few-shot learning classifier under 4 prior conditions:

- **No prior information** To assess the impact of the alignment priors, we compare to a baseline classifier with no prior information (a uniform prior). For this model, SimCLR embeddings for the supervised examples are inputted to the classifier. The classifier optimises performance on categorical cross-entropy loss across the number of classes being tested.

- **Similarity relationship correlations** The similarity relationship correlation does not rely on any trained model to map between spaces. Instead, this prior is based on the similarity relationships to known classes of items, in both image and label spaces. For an image x_i being classified, the similarity relationships to the mean positions of images in seen classes ($x_j \in X_y$ for $y \in Y_s$) are calculated in image space. For each candidate class label $y \in Y_u$, the similarity relationships to known class labels $y \in Y_s$ in the linguistic space are also obtained. Then, the relationships of the candidate image in image space are correlated with the relationships of each candidate class name to the known labels. A prior across candidate class names is generated based on the strength of these correlations. This is visualised in Figure 4.19.

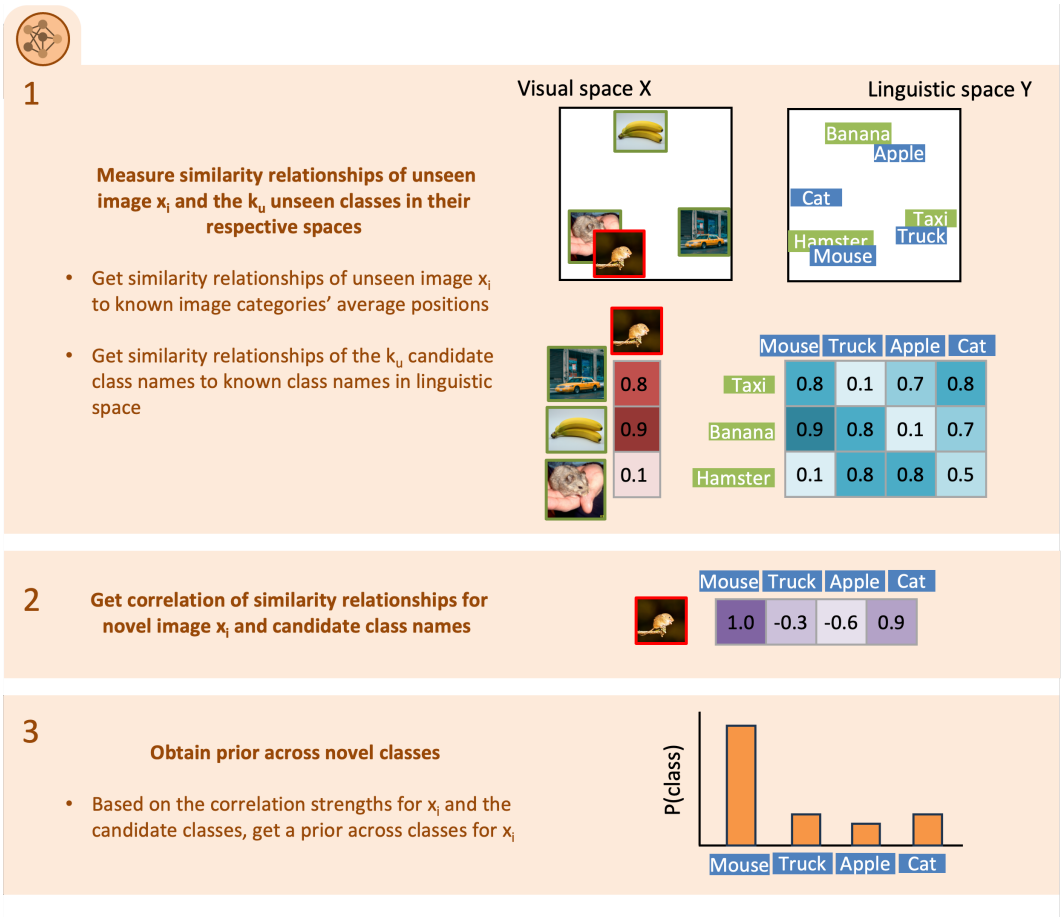


Figure 4.19: Schematic illustrating how the similarity relationship prior is generated across novel class labels based on a set of known mappings for image labels. In box 1, green labels and image outlines represent items from known classes. Blue labels represent the candidate class names, and the red outlined image represents the novel image x_i.

- **Regression model** The regression model learns two mappings between

systems, F(.) and G(.), where F(.) performs the mapping $F : X \rightarrow Y$ from visual space X to linguistic space Y, and G(.) performs the mapping $G : Y \rightarrow X$ from linguistic space Y to visual space X. Both F(.) and G(.) are multi-layer perceptrons, which take inputs in the dimensionality of the source space and output vectors in the dimensionality of the target space. The regression model trains the MLPs using a supervised loss term: for the p% of images in known classes which are used for aligner training, F(.) and G(.) are optimised by backpropagating the L2 distance between the mapping output and the position of the known corresponding item in target space: $\mathcal{L}_{F,Regression} = \frac{1}{b}\sum_{i=0}^{b}(F(x_i) - y_i)^2$ and $\mathcal{L}_{G,Regression} = \frac{1}{b}\sum_{i=0}^{b}(G(y_i) - x_i)^2$, where b is the number of items in the batch. The overall regression loss is the mean of regression losses in each space, $\mathcal{L}_{Regression} = \frac{1}{2}(\mathcal{L}_{F,Regression} + \mathcal{L}_{G,Regression})$.

Once F(.) and G(.) have been trained, the prior across unseen classes Y_u for image x_i is obtained by normalising the inverted distances between $F(x_i)$ and $y \in Y_u$.

- **Cycle + regression model** Much like the Regression model, the Cycle + Regression model also trains F(.) and G(.). The main difference between the regression and the Cycle + Regression models is that the Cycle + Regression model includes an additional component in the loss function: $\mathcal{L} = \mathcal{L}_{F,Regression} + \mathcal{L}_{G,Regression} + \mathcal{L}_{X,Cycle} + \mathcal{L}_{Y,Cycle}$. These cycle loss terms are unsupervised loss components, which aim to minimise the reconstruction loss for mapping x_i back to itself via space Y: $\mathcal{L}_{X,Cycle} = \frac{1}{b}\sum_{i=0}^{b}(x_i - G(F(x_i))^2$, and in parallel $\mathcal{L}_{Y,Cycle} = \frac{1}{b}\sum_{i=0}^{b}(y_i - F(G(y_i))^2$.

  Once F(.) and G(.) have been trained, the prior across unseen classes Y_u for image x_i is obtained by normalising the inverted distances between $F(x_i)$ and $y \in Y_u$.

For each of the three alignment prior conditions (similarity relationship correlations, Regression only and Cycle + Regression), four supervision scenarios are tested: 25%, 50%, 75% and 100% supervision. In the case of the similarity

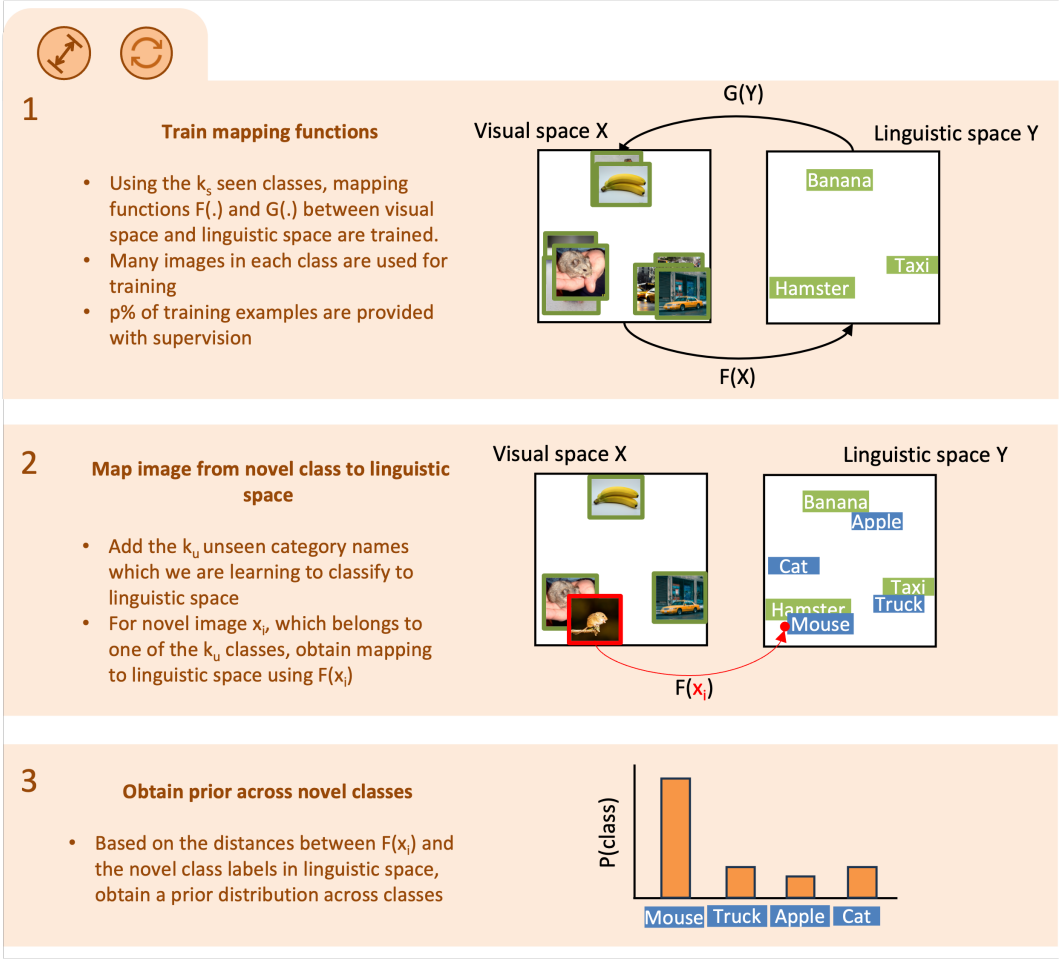


Figure 4.20: Visualisation of how priors across novel classes are extracted from models F(.) and G(.) learned in the Regression and Cycle + Regression models. In box 1, green labels and image outlines represent items from known classes. In box 2, blue labels represent the candidate class names, and the red outlined image represents the novel image x_i.

relationship prior, this means that p% of the images in the known classes are averaged to obtain the mean positions in visual space which are mapped to the known class labels. In the model-based priors - the Regression only and Cycle + Regression priors - this means that the supervised component of the loss term gets contributions from p% of the images in the known classes. For the Regression only prior, the supervised loss term is the only loss term component. The Cycle + Regression model has the additional unsupervised cycle loss terms, which is calculated based on the full set of images in the known image classes.

In each condition, we train 20 classifiers, each on a different sample of known and novel classes. Accordingly, for the model-based priors, a different model is trained for each of the 20 classifiers.

Classification

Following the implementation in the original SimCLR papers (Chen et al., 2020c,b) the classifier used is a simple linear classifier layer with a softmax activation. The classifier's dense layer takes an input of size D_I (the dimensionality of the image space), and outputs a vector of size n_{Y_u} (number of candidate classes).

Where a prior is used, the prior distribution across candidate classes is multiplied by the output of the dense layer. The output of this is normalised to give the final probability distribution across classes. We test models on a classification problem of size $k_u = 100$.

Results

Zero shot performance

A 2-way 3 (prior: Cycle + Regression, Regression only, Correlation) x 4 (supervision: 25%, 50%, 75%, 100%) mixed ANOVA, where prior was a within-subject factor and supervision was a between-subjects factor, revealed a significant main effect of prior ($F(1.64, 122.64) = 1227.29$, $p < .001$) on the uplift in top 1 accuracy from the uniform condition,. No significant main effect of supervision level or the interaction of prior and supervision level was found (full ANOVA table is provided in Appendix C3.1).

Zero-shot performance for all prior conditions is shown in Figure 4.21. As concept sets are nested within supervision levels, post-hoc comparisons are pairwise repeated measures t-tests within supervision levels, with Bonferroni correction applied. In all cases, the Correlation prior performs significantly worse than both Regression and Cycle + Regression models. For some levels of supervision (25% and 75%) the Cycle + Regression model has a statistically significant advantage over the Regression only prior, but this is not consistent across all supervision levels, with a much smaller effect than the difference for the Correlation prior.

Few-shot performance A 3-way 3 (prior: Cycle + Regression, Regression

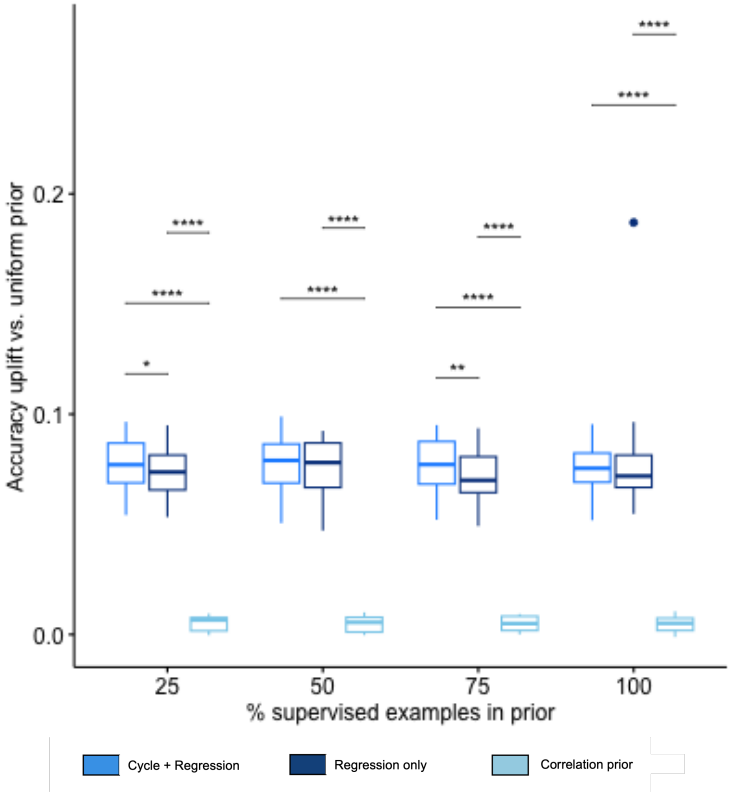


Figure 4.21: Zero-shot classifier performance. Significance in post-hoc pairwise comparisons (via Bonferroni-adjusted paired t-tests) are visualised by *s as follows: *: $p_{adj} < 0.05$, **: $p_{adj} < 0.01$, ***: $p_{adj} < 0.001$, ****:$p_{adj} < 0.0001$. For some levels of supervision, the Cycle + Regression model outperforms the Regression only model, suggesting an advantage for the unsupervised learning mechanism.

only, Correlation) x 4 (supervision: 25%, 50%, 75%, 100%) x 3 (n-shot: 1, 2, 5) mixed ANOVA - where prior and N-shot were within-subject factors and supervision level was a between-subject factor - found significant main effects of prior ($F(1.65, 120.29)=46.90$, $p < .001$) and N-shot ($F(1.83,133.55)=13.90$, $p < 0.001$). No other main effects or interaction terms were found to be significant.

Results for the performances of all prior conditions for 1-, 2- and 5-shot learning are shown in Figure 4.22.

Performance uplift vs the uniform prior condition, with pairwise comparisons by prior condition, are shown in Figure 4.23. In most cases, the Regression and Cycle + Regression models both significantly outperform the Correlation prior but do not perform significantly differently from each other. For the 1- and 2-shot cases at 50% prior supervision, the Cycle + Regression model outperforms the Correlation model but the Regression only prior does not. In both of these cases, the results are trending towards the Cycle + Regression

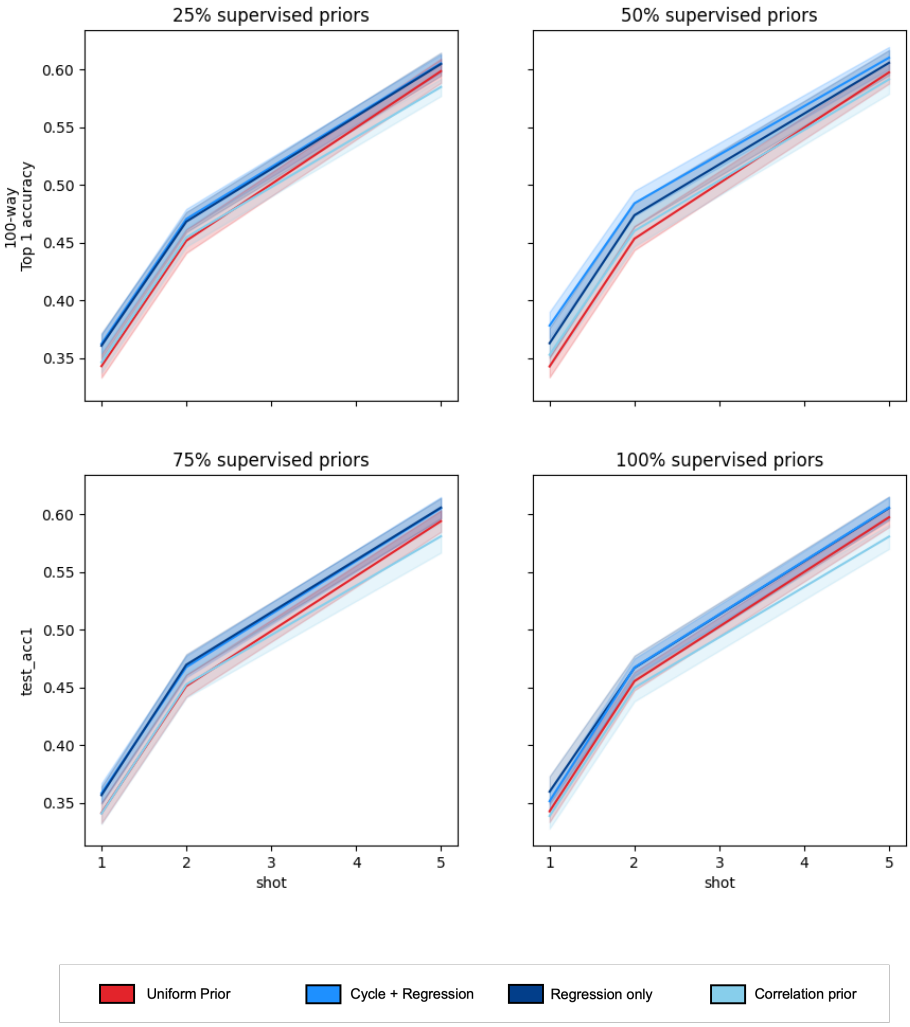


Figure 4.22: Classification performance in few-shot learning compared across prior conditions.

prior outperforming the Regression only prior, but this is not significant at an α level of 0.05.

4.4 Discussion

The first section of this chapter attempted the fully unsupervised version of the cross-modal alignment problem. First, the alignment metric was assessed, by testing modifications to the alignment objective function, in an effort to improve the likelihood of success for models optimising on this score. Modifications to the alignment score were discovered which improved both (a) the correlation between the score and the accuracy of an alignment, and (b) the number of misleading mappings that the score identified (i.e, mappings for

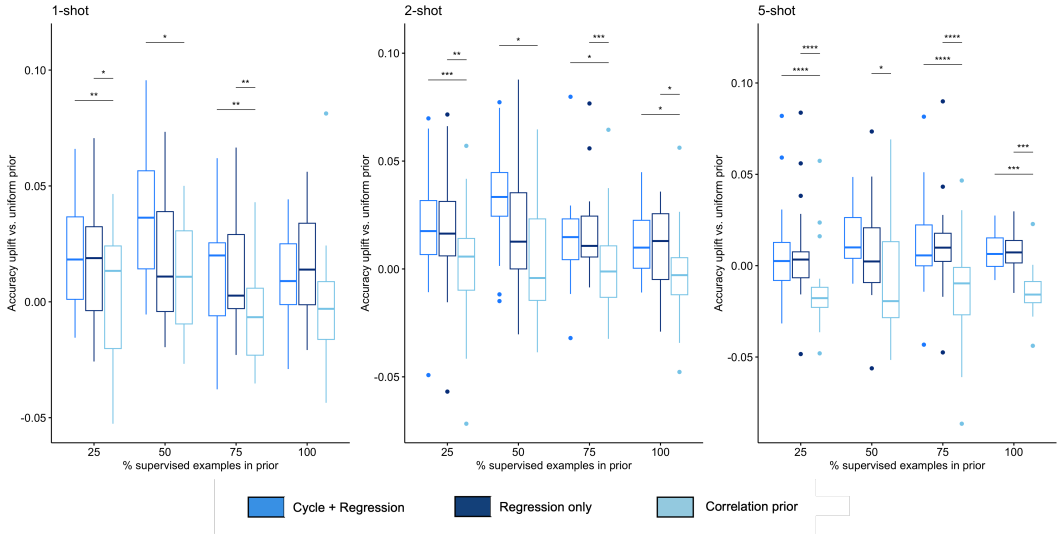


Figure 4.23: Classification performance uplift over uniform prior in few-shot learning. For some levels of supervision below 100%, there is a trend towards the Cycle + Regression model outperforming the Regression only model.

which the score was higher than the score for the correct mapping).

Having identified valuable modifications of the alignment score, a range of algorithms which held promise for solving the unsupervised alignment problem were tested on alignment problems of increasing difficulty. While some algorithms were ideal in noise-free mapping problems, they failed when noise was introduced, and were unsuccessful in the unsupervised mapping problem across modalities. In the fully unsupervised problem, a constrained version of MCTS yielded the greatest success. This algorithm poured computational resources into identifying the most promising matches early in the tree, which led to the highest alignment scores further down the line. This approach was aided by the amendments made to the alignment score. However, while it did improve performance over chance for the larger problem (a 50-concept mapping), the improvements were greatly reduced on account of the heightened computational demands of this search. The results for the smaller problem do provide hope for the future: if MCTS could be implemented at a larger scale, with the heuristics this investigation has identified, performance on larger problems could be improved.

Testing the slate of algorithms on a semi-supervised task, where some concepts were treated as 'known', and new concepts were mapped on this basis, the Cycle model and Kuhn-Munkres algorithm yielded the greatest promise.

The success of the Cycle model was, interestingly, greater than the success a model with the same structure but with no unsupervised loss component included in its training.

These findings led to an investigation of alignment-based priors in image classification tasks. Following the line of work of Socher et al. (2013), Frome et al. (2013) and Akata et al. (2015), priors generated from alignment algorithms were injected into classifiers being trained on novel classes of items in low-data environments. This required the adaptation of alignment systems from a one-to-one mapping problem to a many-to-one mapping problem.

In line with the findings of prior work, Regression and Cycle + Regression priors were effective in yielding performance uplift on the classification task. While the findings of this investigation are somewhat preliminary, there seems to be a trend towards the Cycle + Regression model performing better than the Regression only model, when the models were trained with a certain level of supervision, with this impact appearing to be maximal at 50% supervised training of the values tested.

One reason why this may be the case is that the Cycle + Regression model is pressured to build learn a cross-system mapping which retains the fine-grained local structure of the image embedding space. The cycle loss term aims to reconstruct each image in the original image space via the mapping G(F(x)), and success in this requires the original local similarity structure to be recoverable. With only the regression loss term, as is present in the Regression only model, this local structure is ignored, and the only pressure is to map all images within a class to the class label. This could result in the collapse of the mapped representations of images onto the position of the label. It seems that the presence of images which do not have labels provided for supervised learning are beneficial to the generalisation performance of the Cycle + Regression model, perhaps providing it with the opportunity to optimise for local structure preservation free of the pressures of supervision.

To explore this thought, I performed an analysis of the mappings of training

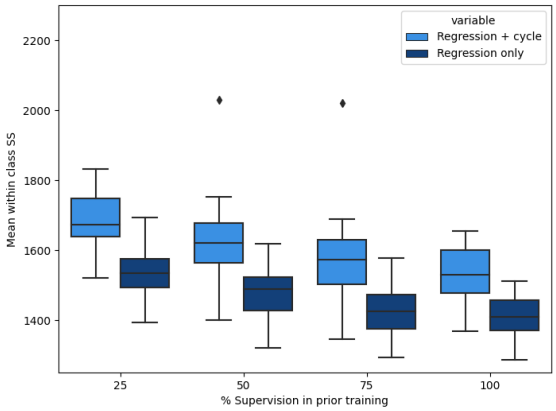


Figure 4.24: The average within-class dispersion for training image classes mapped into linguistic space, compared between Regression and Cycle + Regression models across supervision levels.

images into linguistic space, comparing the the within-class dispersion between regression and Cycle + Regression models. The results are shown in Figure 4.24. These results support this hypothesis, as the within-class dispersion is reliably higher for the Cycle + Regression model than it is for the Regression only model.

This may also give an idea of why the impact of the cycle model is relatively small here, and highlights areas of application where it may be of greater benefit. The image classes in the ILSVRC2012 are, in many cases, highly specific. For example, there are over 100 classes of different dog breeds distinguished in the dataset, from 'kelpie', to 'kuvasz'; from 'flat-coated retriever' to 'curly-coated retriever'. Therefore, there is not much meaningful variation within the visual representations within each class which would be beneficial for generalisation to other classes. But if dogs of all breeds were included under a single label 'dog', a model which retained local structure might be able to generalise more easily to novel classes. The hypothesised difference between models in this example is visualised in Figure 4.25. If generalisation to 'wolf' and 'fox' was tested, a model which retained local structures which was able to disambiguate 'wolf-like' dogs and 'fox-like' dogs may facilitate superior generalisation compared a model whose only aim had been to map all dogs to the label 'dog'.

Based on these findings, a natural next step is to test the effect of a Cycle

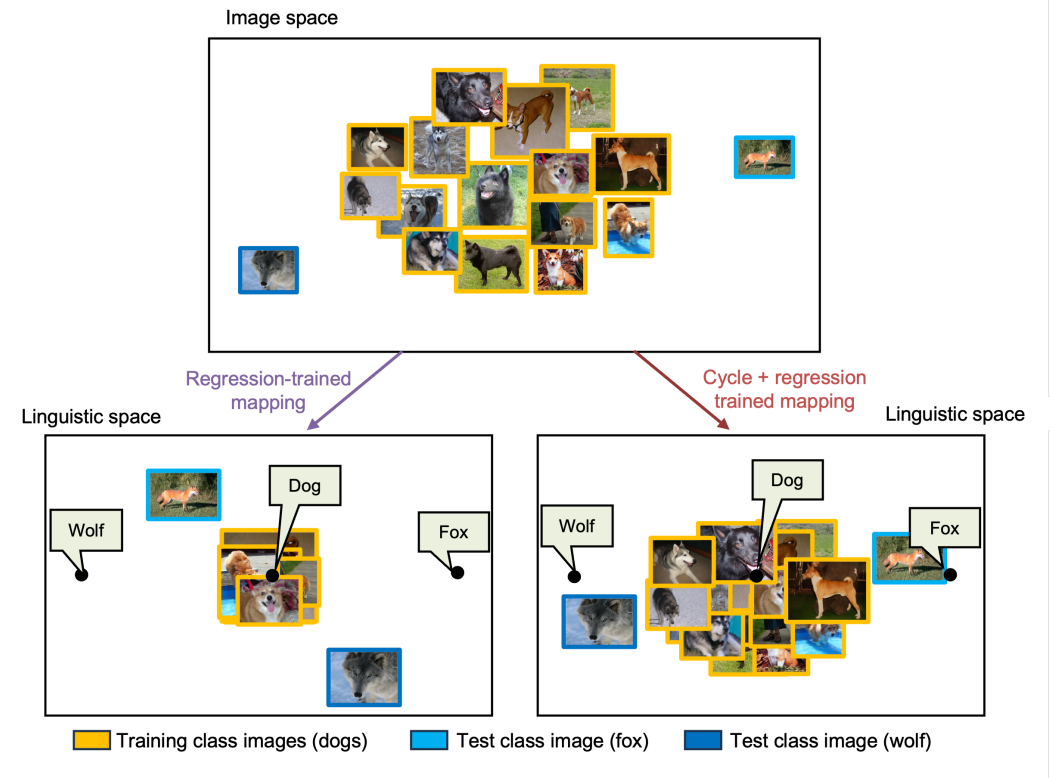


Figure 4.25: Hypothesised local structure preservation in Cycle + Regression model vs label collapse in regression-only models. This Figure demonstrates how the preservation of local structure may assist with successful generalisation to new classes. Here, concept representations both image space and linguistic space are represented by image icons. In linguistic space, the positions of class labels are marked with points and callout boxes. In image space, the organisation of dogs is based on their visual features. When mapped into the linguistic space, the regression model is pressured to map all dog images to the same point in space. The Cycle + Regression model retains the similarity structure in the mapped representations in image space. This allows the Cycle + Regression model to use the local similarity of known classes to generalise to new classes.

+ Regression prior compared to a Regression-only prior for a classification task with more general classes, to see whether the the hypothesised benefit of local structure preservation is indeed more valuable for generalisation in cases where there is more within-class variation. This could be achieved by combining ILSVRC2012 classes at a higher level of the semantic taxonomy.

Overall, this chapter has established a number of promising avenues for machine learning applications of human-inspired alignment mechanisms.

Chapter 5

General discussion

This thesis investigated the role of systems alignment in learning cross-modal mappings.

On the basis of prior work demonstrating that naturalistic information is alignable across modalities (Roads and Love, 2020), Chapter 2 investigated the value of this signal for a challenging real-world cross-modal learning problem: concept learning. By exploring signals for alignment in children's early learning environments, this first study showed that children's early concepts are particularly well positioned to enable learning by alignment. This led to an exploration of the structural features which underpin success in facilitating alignment, guided by insights from the success of children's early concept sets. Having demonstrated the value of alignment signals for learning in the real world, the question remained: do people benefit from systems alignment in learning? In a behavioural study, Chapter 3 tested whether humans do indeed learn from alignment signals when they are available. It showed that alignable systems facilitate more efficient learning, even when full supervision is available. The presence of supervision meant that there was no need to perform alignment to succeed in the task. In fact, for participants in the misaligned condition, performing alignment was counter to the learning goal. And yet, participants in both conditions had a tendency to engage in in systems alignment, which boosted learning performance where systems were alignable. With this finding that alignment facilitates learning in humans, and having

identified the structural signatures of alignment in Chapter 2, potential algorithmic applications of alignment mechanisms were tested. The aim of Chapter 4 was to see whether the psychological and information-theoretic insights from Chapters 2 and 3 could be leveraged to improve the performance of machine systems. Building on the modelling conducted in Chapter 3, a slate of algorithms were tested on the fully unsupervised alignment problem - that is, mapping from linguistic to visual space in the complete absence of labels. In the absence of supervision, only the most computationally demanding algorithms yielded performance advantages over chance. When alignment-based models were used to generate priors for an image classification task (i.e., a cross-modal mapping task), alignment improved classification performance. In some cases, performance was further improved by a alignment models which contained an asynchronous learning mechanism.

In this concluding chapter, I will review the key contributions of the thesis and their significance within the fields of cognitive science and machine learning. I will also discuss limitations of the work presented in this thesis, and highlight promising avenues for future work.

5.1 Could alignment facilitate early concept learning?

The presence of alignable systems across modalities was identified in prior work (Roads and Love, 2020; Riordan and Jones, 2011). In Roads and Love (2020), it was suggested that alignable signals could be beneficial in scenarios requiring learning mappings between systems, for example in mapping visual objects to their labels. Analagous to the naive learners in Quine's 'gavagai' thought experiment, children learning to map language onto the entities they observe in the world are subjected to an overwhelming and very poorly constrained task. Research efforts have long tried to understand how humans are so readily able to learn these kinds of mappings, in spite of the degree of refer-

ential ambiguity inherent therein. But could alignment-based information be providing a valuable constraining signal to the learning problem children face?

Chapter 2 explored this question by simulating real world knowledge development using age-of-acquisition data, and comparing the child-like knowledge states to control knowledge states in their ability to facilitate new concept acquisition via alignment. Regardless of whether linguistic embeddings were pretrained based on large language corpora (Pennington et al., 2014) or trained on smaller, child-directed speech data (MacWhinney, 2000), child-like knowledge states were better than control knowledge states at facilitating learning via alignment. This was true whether the concepts being learned were those that children acquire early in life, or later-acquired concepts.

Alignment has the potential to be a powerful explanatory factor in the 'vocabulary spurt' (Bloom, 2013): alignment postulates that as more cross-modal mappings are known, the easier learning by alignment becomes. Prior to 2 years of age, children exhibit a rapid acceleration in vocabulary growth (Goldfield and Reznick, 1990). This is consistent with findings in alignment whereby the unsupervised acquisition of new mappings becomes easier when more concepts are known.

It must be noted that the causal nature of this finding cannot be discerned from the current exploration alone: do children's caregivers teach them words in early life which provide this superior foundation for future learning from the environment, acting as 'optimal teachers'? Do children learn early acquired words preferentially because they are easier to align, and thus alignment supports the process? The current results do not allow us to parse these possibilities.

One limitation of this work is that, while efforts were made to use child-directed speech embeddings where possible (given the restricted volumes of data), the visual embeddings were not child-directed. While some notable child-like visual data sets from real world environments do exist (Sullivan et al., 2021; Clerkin and Smith, 2022), to date none allow for the straightforward ex-

traction of representative child-like visual embeddings. I would hypothesise that the alignment of child-like visual embeddings and child-like linguistic embeddings would be even stronger than those observed in the current study, as both would be subject to the same environmental constraints as children are in the world.

A crucial direction for future work here is to test the extent of learning by alignment in children, via behavioural studies. While this exploration demonstrates that the signals in children's early concepts support learning by alignment, it does not conclusively demonstrate that they use these signals in learning. Given the signal identified in this work and the finding that adults learn from alignable signals, the clear hypothesis is that the same would be found in children.

Based on the finding that early-acquired concepts yielded a greater capacity for alignment based learning than control concept sets, the next question in the investigation presented in Chapter 2 was: are there structural features which make a concept set particularly well-suited for alignment-based learning? The distinguishing structural features of child-like concept sets, which had performed well in the alignment learning task, were analysed to address this question. It was shown that there were indeed distinguishing structural features at play. Specifically, early knowledge states had dense connectivity, both within the knowledge state and with concepts outside of the knowledge state (i.e., knowledge yet to be acquired). They also contained concepts with more proximal nearest neighbours, had lower dimensional coverage, and had more positively skewed degree distributions.

Noting that these concept sets conflated being early-acquired and yielding success in structural alignment, the influence of structural features on alignment was tested using generative agent-based modelling. Agents' internal parameters controlled the structural features of interest identified from the child-like knowledge states. One class of agent optimised these internal parameters such that the knowledge states it produced were as similar as pos-

sible to the child-like knowledge states produced using age-of-acquisition data; the other class of agent optimised its internal parameters such that it maximised performance in the alignment-based forced choice task. For the agent which matched to Age-of-Acquisition (AoA) data, the learned internal parameters matched well with the parameters identified in the analysis, with the exception of the most proximal nearest neighbour, which was not influential in deciding the concepts to select. For the agent optimising task performance, the key structural factors in concept selection were maximising connectivity, both within and beyond the knowledge state, and high degree skew within the knowledge state.

While the Task-Optimised agents did outperform the AoA agents in alignment-based forced choice, the results showed that the structural features of the AoA concept set were close to optimal for alignment-based learning, as the difference in performance between early-acquired concept sets and task-optimised agents was relatively small. The success of the Task-Optimised agent is attributable to its focus on dense connectivity in the concept space, while the AoA-Matched agent had to prioritise other features in order to match the statistics of early concept acquisition, which were non-essential for alignment but perhaps key for other aspects of conceptual development (Hills et al., 2009; Stella et al., 2017).

The gap in performance between the AoA-Matched agents and the AoA agents suggests that perhaps something is missing from the current structural feature set, which underpins the success of the AoA agents in alignment-based forced choice in these early months. One explanation for this is that the internal model parameters could be dynamic, and evolve over time, which is not accounted for in the model presented here. For example, the dense connectivity of early-acquired concepts could be key for the concepts acquired in the first months, and this could be reflected in the structural features of early concepts in the real world. But after this initial knowledge base is built, the concept acquisition strategy in infancy could evolve, for example to deepen

knowledge within sub-categories, and this could change the balance of structural features' influence. Future work could look to incorporate dynamics into the internal models. This has the potential to improve performance for both the AoA-matching and the task-optimised models, and thus to develop our understanding of how alignment might be supported in early concept learning.

This structural analysis identified structural underpinnings of alignment, and the generative modelling portion successfully applied insights from the psychological world to improve the performance of machine-learning systems in this task.

Having demonstrated that alignment signatures exist in a naturalistic learning context, the investigation moved into the lab, where humans' use of alignment signals for learning was tested and modelled.

5.2 Does alignment support human learning?

Having explored the value of alignment signals for learning in the real world, Chapter 3 presented a controlled behavioural study in the lab, to test whether humans benefit from alignable signals when learning to map between systems. This was a novel investigation of whether underlying second-order isomorphism improved human performance in a learning task. Performing this experiment in the lab allowed models to be fit to each individual participant's behaviour, in order to explore the best fitting learning mechanisms across conditions.

The hypothesis of the paired-associate learning experiment presented in Chapter 3 was that learning would be more successful when the correct correspondences were dictated by systems alignment (aligned condition), than when the correct correspondences were randomly selected (misaligned condition). Participants were not instructed to align in the learning task, providing a good test of whether people have a tendency to perform alignment based on the underlying structures of systems. The results provided support for this hypothesis: participants in the aligned condition performed better in the learning task than those in the misaligned condition. Furthermore, participants in the

aligned condition were able to successfully generalise their cross-system mapping to a completely novel example - or in other words, to perform zero-shot learning.

In both conditions, participants were given full and repeated supervision: they were shown the correct responses twice before each block of trials. As such, there was no need to engage in systems alignment to successfully perform the task at hand. What's more, computational modelling of each participant's series of responses found that a model which contained an unsupervised alignment mechanism was the best fit for the majority of participants - not only in the aligned condition, but also in the misaligned condition. The unsupervised alignment mechanism attempted to map the full set of representations in one domain onto the set of representations in the other. As the correct responses for participants in the misaligned condition were not consistent with this kind of smooth mapping, the alignment mechanism did not benefit the performance of these participants. And yet, the errors that participants in this condition made were consistent with a tendency to align systems.

The alignment mechanism in the best fitting model of human behaviour attempted to align the two systems as wholes, rather than just learning a mapping from one space to another. This model, referred to as Cycle + Regression, and inspired by Zhu et al. (2017)'s CycleGAN, operated with an unsupervised loss term. While in the models fitted here, alignment steps were interleaved with supervised trials, this optimisation step could be performed asynchronously in the absence of any supervision at all, based on the distributions of items in each system independently. This could, for example, occur via neural replay (Barry and Love, 2023). Since the study in Chapter 3 was conducted, an EEG study conducted by Huang and Luo has demonstrated that replay contributes to improving working memory for aligned systems: spontaneous replay of one system occurred when a second system was recalled, only when the two systems are aligned (Huang and Luo, 2023). This provides strong support for the hypothesis that replay could facilitate cross-modal learning

from signals presented asynchronously in the real world (Clerkin and Smith, 2022; Tamis-LeMonda et al., 2019).

As a whole, the research project oriented around alignment-based signals in learning is firmly rooted in the goal of understanding learning beyond the lab: alignment-based signals have been identified in naturalistic environments, and as such represent a rich and complex information source that is available in the real world. Chapter 3 of this thesis is the first behavioural study exploring the effects of systems alignment in cross-domain learning, and demonstrated the hypothesised effect. But one clear next step would be to explore alignment effects in a more naturalistic setting: the artificial stimuli and controlled experimental paradigm limit the scope of the conclusions that can be drawn about the role of alignment in the real world at this stage. While the finding that learning is supported by alignable systems is certainly novel, the nature of learning by alignment outside the lab remains to be explored. One step in this direction would be to update the stimulus space to be more complex or of higher dimensionality, for example by using semantic dimensions as the axes of variation, and selecting visual concepts as the stimuli.

Extensions of this work could examine how the alignment effect changes with varying degrees of task supervision. In Chapter 3, the supervision signals were complete: participants were shown the correct mapping for every item. Testing the alignment effect in a weakly-supervised context could provide further insight on how the effect may interact with other learning signals in the real world, as supervisory signals are rarely so clear in reality. It is possible that weakly supervised contexts would lead to a larger alignment effect than was observed here: alignment-based learning is theoretically possible with no supervision at all, but the same is not true the misaligned condition. So, reducing the supervision to somewhere between the perfect supervision here and no supervision is likely to yield benefits for the alignment condition.

Another factor of interest is the size of the learning task (i.e., the number of items to learn). In this study, the task was carefully formulated such that

success was attainable for the vast majority of participants, in either condition. But it is feasible that more difficult tasks would demonstrate even more of a performance gain from alignment. There may be a trade-off between the cognitive costs of aligning systems and memorising individual mappings. For example, if the number of items to learn mappings for was far too high to retain in working memory, alignment could reduce the memory requirements of the task sufficiently for a substantial difference in success rates (Edelman, 1998).

5.3 Can machine learning systems learn by alignment?

The power of alignment in human learning exhibited in Chapter 3 of this thesis, along with the finding that children's early concepts are near optimal from an alignment perspective, begs the question of whether alignment-based mechanisms can benefit machine-learning systems in learning contexts where humans excel and machine systems struggle.

The most challenging formulation of this question entailed exploring whether fully unsupervised alignment across modalities was possible using alignment. That is, can the similarity structures within individual modalities be mapped across modalities with no known mappings to start with? This is a very difficult problem at scale: the size of the candidate mapping space makes search intractable, and the distinct morphologies within the relevant systems mean that tricks used in other applications of alignment do not apply (e.g., anchoring the alignment with proper nouns, which share their morphology across multiple languages, in language alignment for translation, Conneau et al., 2017). Solving the problem of unsupervised alignment across modalities would be a huge breakthrough in Machine Learning. Chapter 4 of this thesis presented an exploration of potential avenues for solving this problem at scale. Given the difficulty of the unsupervised alignment problem, applications of an unsu-

pervised alignment mechanism to semi-supervised cross-modal learning problems was also explored, motivated by the findings of previous chapters which showed that (a) humans use alignment signals alongside supervision to improve learning, and (b) alignment in tandem with prior knowledge can facilitate the acquisition of new mappings.

Building on the modelling in Chapter 3, which modelled alignment processes in conjunction with supervision, Chapter 4 scaled up the alignment problem, and dialled up the difficulty by testing the alignment algorithms where no supervision at all was available. A range of unsupervised methods, which have demonstrated success on comparable problems through different lenses, were tested.

The investigation first looked to set the algorithms up for maximum success by testing modifications to the objective function that algorithms sought to optimise. Modifications covered dimensionality reduction of the unimodal representations, transformations of similarity relationships and different correlation functions for scoring the correspondence of similarity relationships. While some of these modifications brought about improvements in the correspondence between score and mapping accuracy, these improvements were not reflected in improvements in algorithm performance. Monte Carlo Tree Search algorithms showed promise in identifying successful mappings across systems, but versions which were most successful were highly computationally intensive. Future work could build on the successes of this algorithm by using more efficient MCTS modifications, to explore the potential for MCTS methods to succeed in unsupervised cross-modal mapping problems.

On the introduction of some supervision, in the form of some known mappings between systems, alignment success also begins to improve for other models, as we would expect from the prior work of Socher et al. (2013), Frome et al. (2013) and Akata et al. (2015). Interestingly, preliminary findings in this thesis suggest that an unsupervised alignment component when learning the cross-modal mapping from known concepts could lead to improvements in

generalisation to new classes of items, compared to mapping functions trained with a purely supervised signal.

Trends in this direction were found when comparing the impact of priors in few-shot image classification. Priors were generated using alignment-based models. Mapping functions were trained on a set of seen image classes. Once trained, these mapping functions were used to generate priors for images in unseen classes. While priors improved classification performance in general, it was found that the presence of unsupervised examples and an asynchronous alignment mechanism gave the best prior of all. When there were unsupervised examples to learn from, mapping models which were trained with a cycle loss term - the same model which provided the best fit to human learning in Chapter 3 of this thesis - were better able to generalise to previously unseen image classes than either (a) models with the same amount of supervision, but no cycle loss term, or (b) models including the cycle loss term, but with supervision for all training examples.

Why would unsupervised examples in mapping function training yield superior classification performance for unseen classes of items? I presented evidence that models trained with some unsupervised items may develop a greater appreciation for local similarity structure. It seems that the presence of some items whose label is not specified, and which the model is therefore not pressured to map onto the specific location of the label in label space, is beneficial in facilitating model generalisation. Future work on different datasets, perhaps with more general labelling schema, could explore this effect further, to understand the potential value of incorporating unsupervised alignment into model training. In situations where labelled data is rare, this has the potential to extract more information from unlabelled data, and thus could be valuable in machine-learning domains where obtaining labelled data is a challenging roadblock.

5.4 General limitations

The limitations of the individual components of this thesis have been discussed in the prior sections of this chapter. In this section, I address limitations to more general aspects of the approach taken in this thesis.

One general limitation to note is that modelling cannot in isolation prove the cognitive processes at hand. No matter how strong a model fit may be, it will never be able to provide conclusive evidence that it captures the processes taking place in the mind (McClelland, 2009). The models used throughout this thesis serve to explore alignment as a theoretical framework, and to understand the potential implications of alignment-based processes for learning in human and machine-systems. So, this limitation does not detract from the value of computational models of cognitive processes, which allow us to formalise theory and generate testable predictions. Instead, the computational work presented in this thesis paves the way for future studies to test the mechanisms underpinning the alignment-learning effect in humans, observed in the study presented in Chapter 3.

Throughout this work, embeddings are used as approximations of unimodal information available to humans. In reality, humans' experiences develop over time and are not static like our embeddings. In applications for early concept development, this raises interesting questions about the correspndences between developing unimodal spaces. In terms of the inputs to modalities, it seems highly plausible that the constraints on children's early environments constrain multiple modalities in similar ways, and thus may not impede alignment processes. This is not to say that the inputs to different modalities are perfectly matched to one another - indeed, we have discussed evidence throughout this thesis that synchronous inputs relatively rare (Clerkin and Smith, 2022; Tamis-LeMonda et al., 2019) - but merely that most infants experience a narrower range of environments than most adults, and that the language they are exposed to most frequently is reflective of this (Tamis-LeMonda et al., 2019). So, from an input perspective, alignment could operate on (and

even benefit from) gradually developing unimodal spaces, as the number of concepts for which representations exist to be mapped reduce the size of the problem space.

Embeddings also possess a degree of randomness. This was demonstrated in Chapter 2, where multiple initialisations of an embedding algorithm yielded differences in similarity structure (particularly for long-range relationships which, I have argued, carry less meaning). This is perhaps analogous to differences in the representations humans infer from their idiosyncratic experiences of concepts throughout their lifetimes, which may lead to differences in how inter-concept relationships are structured across individuals. However, the nature and extent of this variability has not been compared to variability in human representations. The findings in this thesis persist in spite of the potential for variability in the inferred embeddings, generalising across adult- and child-like embeddings in Chapter 2 and from co-occurrence based visual embeddings to visual similarity-based embeddings in Chapter 4. Nonetheless, it is worth acknowledging that the relationship between variability in embeddings and human concept representations is speculative, and therefore noting that variability in human unimodal representations may have different impacts on alignment than those suggested by computational models.

5.5 Potential implications of alignment

The theoretical perspectives explored in this thesis provide a new view on how learning can occur from information in different modalities. Alignment signals were shown to benefit machine learning across all chapters, and to benefit human learning in Chapter 3. Developments in the theory of learning are valuable in their primary contribution to our understanding of how the human mind works, and establish directions for research connecting these cognitive findings to their neural underpinnings. As was demonstrated in this thesis, theoretical developments in learning - such as the finding that humans can learn by alignment - are also valuable in their application to machine learn-

ing systems. In Chapter 2, I demonstrated that agents were able to develop knowledge states which were optimal for unsupervised generalisation to new concepts, based on the structural features of early-acquired concepts. In Chapter 4, the application of unsupervised alignment mechanisms which captured human performance in an alignment task have shown potential in application to cross-modal machine learning tasks. Future applications to the domain of machine learning could include the development of cross-modal concept curricula for machine learning systems (Bengio et al., 2009), with the aim of setting networks up for success in generalisation to new image and word concepts focusing on alignable concepts in their early training (Elman, 1993; Lake et al., 2017).

Outside of machine learning, the perspective on learning offered by an alignment account could be beneficial in developing learning strategies which may be useful when the integration of multimodal information into concept representations is impaired. In semantic dementia (SD) patients, the deterioration of semantic memory results in deficits in concept understanding, speaking and object recognition (Warrington, 1975; Hodges et al., 1995), while other cognitive functions such as episodic memory and topographic memory remain largely unimpaired (Pengas et al., 2010). Aiming to improve the quality of life of SD patients, research has aimed to develop strategies that patients can use to retain their conceptual language (i.e., to slow progressive anomia), and even to recover or build their semantic knowledge. These strategies are often called 'relearning techniques'. Some such strategies include straightforwardly repeating exposures to object pictures and names, while other approaches involve linking semantic concepts to objects used in the patients own home or environment (Robinson et al., 2009), and linking concepts to their autobiographical memories (Snowden and Neary, 2002). Many of these strategies yield some success, but struggle to help patients generalise beyond the learning context (e.g. the specific task or exemplars used). Suárez-González et al. (2015) showed that generalisation of restored concepts beyond specific exem-

plars was substantially improved by using a 'cognitive enrichment' relearning strategy which linked degraded concepts to the patient's remaining semantic knowledge.

Given that alignment has been found to benefit cross-modal mappings (such as those required for visual object naming), alignment-based strategies could be tested in similar contexts to see if semantic dementia patients benefit from these signals in maintaining their semantic networks. For example, by presenting training examples within naturalistic visual and linguistic contexts, generalisation to new examples in the real world may be bolstered by alignment processes.

Along similar lines, alignment strategies could be tested for children with word-learning difficulties. Nash and Donaldson (2005) demonstrate that children with word learning difficulties struggle to learn at the same level as children with normal word learning in both incidental and explicit learning contexts. It is possible that alignment could benefit these children, for example by guiding their attention in incidental learning environments to the information contained within the context of a word. Alternatively, using principles like those explored in the agent-based modelling study in Chapter 2, alignment principles could develop curricula for word learning, to help these children build knowledge bases which facilitated the strongest possibility for alignment-based learning when exposed to new and unknown words.

These potential applications, together with the findings of the work presented in this thesis, identify a range of exciting avenues for future work.

5.6 Future directions

The discussion above has identified several directions for future pursuits in the application of systems alignment. Some of these involve the use of alignment to improve long-term memory of cross-system mappings. In the context of memory, alignment could be viewed as an abstract schema (Bartlett and Bartlett, 1995) - a conclusion supported by the finding in this thesis that peo-

ple have a tendency to align items, even when this is not beneficial to the task. This finding suggests that humans expect systems to align. Prior work on schemas suggests that memory for schema-congruent and schema-incongruent items are both improved relative to schema-irrelevant items, and that these two processes are mediated by different regions of the brain (Van Kesteren et al., 2012; Greve et al., 2019). Future work should investigate the impact of alignment and misalignment across systems on memory for cross-system mappings, perhaps testing the hypothesis that alignment operates as a schema for cross-modal mapping. If memory under alignment was indeed improved, this could guide efforts to use alignment in order to address the memory difficulties discussed in the previous section.

Above, I presented many applications of systems alignment for mapping between visual and linguistic spaces, from further behavioural tests to probe alignment in human learning – in both adults and children – to extensions of the machine learning approaches tested, in pursuit of a solution to the unsupervised visual-linguistic mapping problem. But systems alignment is likely a general learning process. While much of this thesis explored the implications of alignment in visual-linguistic cross-modal learning (based on prior work showing that alignable signals exist for this task in the naturalistic environment) the findings of the Chapter 3 suggest that alignment could apply to other domains. The behavioural study presented in this chapter demonstrated that alignment plays a role in a general cross-system mapping problem, and that learning in this scenario is well explained by a model which contains an unsupervised alignment mechanism. As such, future applications of alignment beyond the visuo-linguistic realm.

Mappings exist between many kinds of systems. Consider the correspondences between sounds and shapes exhibited by the 'bouba'-'kiki' effect (Köhler, 1970; McCormick et al., 2015); between music and emotions (Juslin et al., 2001; Eerola and Vuoskoski, 2012) and (perhaps obviously) between languages. Indeed, perhaps because some valuable alignments do exist, people also seek

alignments elsewhere - for example, by mapping birthdays to personality traits via astrology.

In translation, alignment may provide a lens through which to examine word meaning. By identifying translations for which alignment is poor, one could explore cross-linguistic mappings which do not do a good job of capturing meaning in translation. It may also prove to be a useful tool for analysing translation errors. It is also possible that, by examining changes in a word's position within the linguistic space over time, we could use alignment to understand changes in word meaning within a single language (Hamilton et al., 2016), which could have implications for our understanding of historical documents.

In the example of music-emotion mappings, Won et al. (2021) took an approach similar to the cross-modal approaches used in image retrieval tasks to retrieve music based on emotional language. It is possible that alignments between physiological response spaces and music-spaces could add further to a multimodal representation of emotional musicality.

When it comes to learning by alignment in these broader domains – particularly where belief is involved, such as in the astrology example – issues of motivation may prove important. People's representations of the world, the information they seek out within it, and the extent to which information induces learning, are all impacted by individuals' motivations and existing beliefs (Leong et al., 2019; Sharot and Sunstein, 2020; Kappes and Sharot, 2019). In the context of alignment, such factors may influence the extent to which alignment-based learning occurs, or may influence the representations of the systems being aligned themselves. To understand learning by alignment in a broader range of contexts, the value of alignment for human learning under different motivational conditions could be explored.

Alignment as a framework has the potential to help us understand cross-system mappings and their implications beyond the realm of cross-modal learning. If future work continues to show that humans apply alignment to learning

in a range of domains, as the work in this thesis hypothesises, this can be incorporated into machine systems to build human-like mechanisms of information processing.

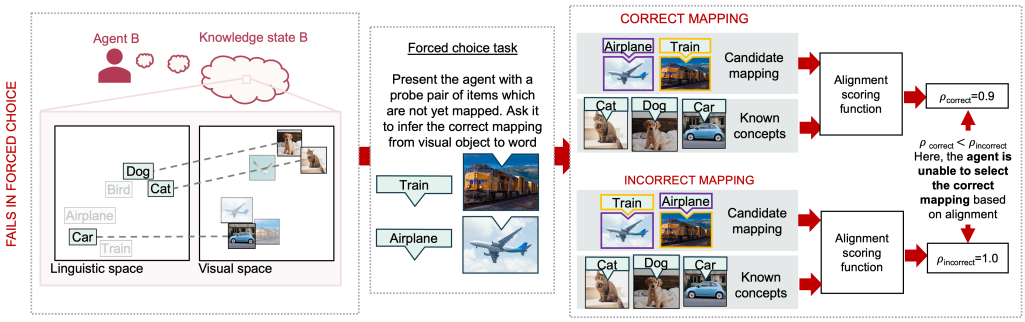
5.7 Conclusion

This thesis has explored cross-modal alignment in learning, shedding light on the oft-neglected unsupervised learning processes which facilitate the integration of sensory inputs into meaningful representations of the environment. It has demonstrated that alignment signals are valuable in the naturalistic cross-modal learning problem of concept acquisition. Together with its finding that humans learn and generalise better across systems where their underlying structure is constrained by alignment, this presents a revised account of learning, which incorporates asynchronous processes that capitalise on emergent cross-modal structure. Computational modelling of human learning identified promising candidate mechanisms for these asynchronous processes, which were applied in machine-learning solutions to unsupervised and zero-shot generalisation problems. Preliminary successes here open the door for alignment methods to capitalise on unlabelled data in cross-modal tasks. These human-inspired models take steps towards addressing widespread problems with the availability of labelled data, by emulating the human capacity to learn by asynchronous alignment. I hope this thesis offers an exciting new perspective on the role unsupervised learning processes could play in human and machine learning in the real world.

Appendix A

Supplementary information for Chapter 2

A1 Example of failed alignment



An example of an Agent failing the forced choice task based on its starting knowledge state.

A2 Forced-choice experiment with CHILDES

The results for the forced-choice experiment conducted using the resultant child-directed speech embeddings are shown in Figure A.2, with ANOVA results provided in Table A.1. The results of the experiment echo the results when the pre-trained GloVe embeddings were used: we find a significant main effect of agent type ($F(1, 198) = 1165.70, p < .001, \eta_p^2 = 0.850$) and probe type ($F(1, 198) = 12.03, p < .001, \eta_p^2 = 0.070$), as well as a significant agent × probe type interaction ($F(1, 198) = 6.90, p = .009, \eta_p^2 = 0.049$).

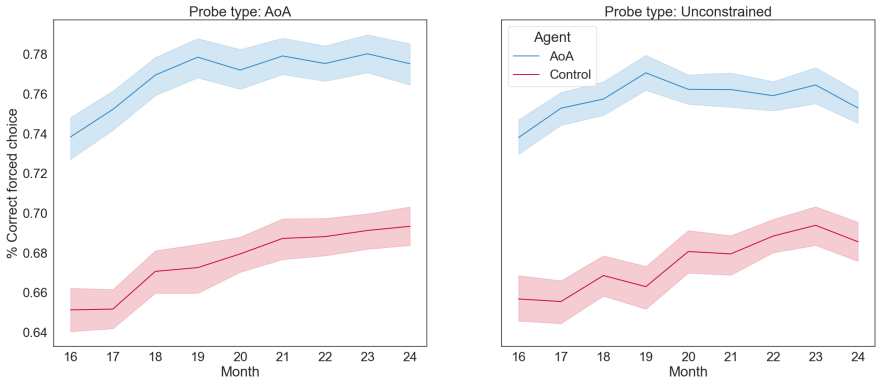


Figure A.2: Forced choice results where linguistic embeddings are derived from child-directed speech corpus CHILDES.

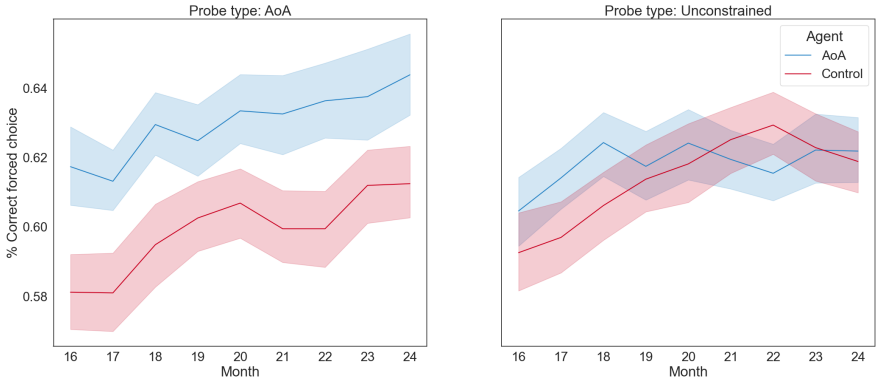


Figure A.3: Forced choice results where linguistic embeddings are derived from the enwik8 corpus, downsampled to match the CHILDES corpus in size.

Predictor	df	F	p	η_p^2
Agent	$(1, 198)$	1165.70	$< .001^*$	0.850
Probe	$(1, 198)$	12.03	$< .001^*$	0.070
Agent * Probe	$(1, 198)$	6.90	$.009^*$	0.049
Month	$(8, 1584)$	25.50	$< .001^*$	0.111
Agent * Month	$(8, 1584)$	4.55	$< .001^*$	0.020
Probe * Month	$(8, 1584)$	1.65	0.107	0.007
Probe * Agent * Month	$(8, 1584)$	0.49	0.863	0.002

Table A.1: Repeated-measures ANOVA results for probe pair experiment with word embeddings derived from CHILDES dataset of child-directed speech. Agent condition (AoA vs. control) was a between-subject factor, and probe condition (AoA-constrained vs. Unconstrained) and month were within-subject factors. $*$ indicates statistically significant results for α=0.05. df = degrees of freedom; η_p^2 is partial η^2 effect size.

Embeddings derived from a sample corpus drawn from enwik8, of comparable size to the CHILDES corpus, did not yield the same results. The results for this corpus are shown in Figure A.3, and ANOVA results are provided in Table A.2.

Predictor	df	F	p	η_p^2
Agent	$(1, 198)$	49.27	$< .001^*$	0.200
Probe	$(1, 198)$	0.31	0.577	0.015
Agent * Probe	$(1, 198)$	17.46	$< .001^*$	0.092
Month	$(8, 1584)$	112.98	$< .001^*$	0.059
Agent * Month	$(8, 1584)$	1.45	0.169	0.006
Probe * Month	$(8, 1584)$	1.10	0.364	0.005
Probe * Agent * Month	$(8, 1584)$	1.04	0.401	0.004

Table A.2: Repeated-measures ANOVA results for probe pair experiment with word embeddings derived from the enwik8 Wikipedia dataset. Agent condition (AoA vs. control) was a between-subject factor, and probe condition (AoA-constrained vs. Unconstrained) and month were within-subject factors. $*$ indicates statistically significant results for α=0.05. df = degrees of freedom; η_p^2 is partial η^2 effect size.

A3 Pairwise t-tests vs. chance for Control and AoA agents

	Control		AoA	
Month (m)	t-statistic	p	t-statistic	p
16	88.86	$\ll 0.001^*$	180.27	$\ll 0.001^*$
17	88.69	$\ll 0.001^*$	167.68	$\ll 0.001^*$
18	105.51	$\ll 0.001^*$	194.75	$\ll 0.001^*$
19	98.09	$\ll 0.001^*$	190.28	$\ll 0.001^*$
20	111.38	$\ll 0.001^*$	198.06	$\ll 0.001^*$
21	101.42	$\ll 0.001^*$	202.29	$\ll 0.001^*$
22	102.06	$\ll 0.001^*$	207.49	$\ll 0.001^*$
23	114.62	$\ll 0.001^*$	194.81	$\ll 0.001^*$
24	113.06	$\ll 0.001^*$	189.24	$\ll 0.001^*$

Table A.3: One sample t-test results for the comparison of control and AoA forced-choice results to chance performance (50% accuracy) with pre-trained embeddings. At $\alpha = 0.05$, Bonferroni corrected for 18 individual comparisons to give adjusted threshold 0.0027, all comparisons are highly significantly different from chance performance in the forced choice task.

A4 Features tested

Feature type	Feature	Description
Node	Distance in full space (mean/-max/min)	The mean/max/min magnitude of a concept's distances from other concepts in the full concept space.
	Distance within knowledge state (mean/max/min)	The mean/max/min magnitude of a concept's distances from concepts in the existing knowledge state
	Degree in full space (k_{full})	Number of vertices between a concept and all other concepts in the full space.
	Degree in knowledge state ($k_{\text{knowledge}}$)	Number of vertices between a concept and all concepts in the existing knowledge state.
	Betweenness in full space	Fraction of shortest paths in the full space graph which pass through the concept.
	Betweenness in knowledge state	Fraction of shortest paths in the knowledge state graph which pass through the concept.
	Clustering in full space	The fraction of possible triangles that pass through the concept in the full space which are realised.
	Clustering in knowledge state	The fraction of possible triangles that pass through the concept in the knowledge state which are realised.
Knowledge state	Average dimension coverage	The average proportion of embedding dimensions' overall variability in the full concept set which is covered by the concepts in the knowledge state.
	Degree distribution skew (proxy)	The skew of the degree distribution of the knowledge state. Skew is approximated as $\frac{\max(k)-\text{mean}(k)}{\max(k)-\min(k)}$

Table A.4: Features tested for knowledge state classification

A5 Bootstrapping AoA distributions for AoA-matched loss

Generating bootstrapped AoA distributions is necessary so that we have training and validation data for model selection across restarts. It also builds in an acknowledgment of the fact that the WordBank dataset is itself a single sample from the underlying population distribution of word acquisition. The bootstrapping process is as follows:

- We generate $B = 1000$ bootstrapped probability distributions, $\tilde{P}_b(X)$.
- For each one, we sample 300 AoA sequences using the procedure described above and visualised in Figure 2.7. Then we calculate $\tilde{P}_b(X)$ from these generated sequences, by calculating the proportion of sequences in which each concept was acquired by each month.

- Generate train and validation sets of bootstrapped distributions (70/30 split for train/test).

As the magnitude of the loss varied by month, we aimed to normalise the loss month-wise such that no month was disproportionately favoured in optimisation. To achieve this, we sampled 5,000 MSEs for each month m, by randomly selecting N_m concepts and calculating the average MSE between the resultant probability vector and bootstrapped probability distributions. We then calculated a z-score for the loss term using the mean, μ, and standard deviation, σ, of MSEs acquired for the relevant month. To ensure that no loss was below zero, we subtracted the z-score of the theoretical minimum MSE (0) from all MSE z-scores. This constituted the final loss term for the AoA-Matched agent.

A6 Soft alignment loss for Task-Optimised agents

The calculation of the soft alignment loss proceeds as follows:

- For the optimal agent, the loss requires a sample of test pairs for the forced choice experiment. Therefore, on each backpropagation step, we segment the remaining concepts into the 'candidate set', the 'train set' and the 'validation set'. The candidate set contains 300 concepts, and is comprised of the n_t concepts in the current simulated knowledge state, and $300 - n_t$ concepts randomly selected from the remaining concepts.

- This leaves 59 concepts for each of the training and validation concept sets, from each of which 750 random pairs of concepts are then sampled to serve as the testing and validation slates respectively.

- As in the training procedure for the structural agent, at each timestep we obtain a vector across all candidate concepts, whose value represents each concept's probability of being in the knowledge state at $t+1$. Therefore,

for concepts which have already been acquired, the value of this vector is 1; for any concepts which have not yet been acquired, the value is determined by the probability obtained from the generative score.

- This is the same probability vector used in the structural agent's training process. In the case of the optimal agent, this probability distribution is used to weight the contributions of inter-concept distances to the alignment score.

- We obtain the pairwise probability matrix $\mathbf{p}^T\mathbf{p}$, and use this to weight the Spearman correlation in the alignment score calculation for the two permutations of the test pair mapping.

- At each timestep, we backpropagate this soft alignment loss to update the target variable vector $\mathbf{x}$ and the weight vector $\mathbf{w}$.

Following the observation that the magnitude of the soft alignment loss increases with the number of concepts in the knowledge state, we normalised the alignment loss terms within each month. To achieve this, we aimed to replicate the alignment loss calculation as closely as possible:

- For each of 5,000 samples for each timestep t, we sampled a pseudo-partition of 300 concepts, from which we sampled a knowledge state of size $n_t - 1$.

- We then gave all of the selected concepts a probability of 1, and generated a uniform distribution across the remaining concepts in the pseudo-partition, just as is done prior to the soft alignment score calculation in training.

- We then calculated the difference between the correct and incorrect alignment scores for a randomly selected pair of concepts from outside of the partition set for each sample ($\mathbf{s}_{\text{incorrect}} - \mathbf{s}_{\text{correct}}$).

To normalise the loss terms in model training using the distribution of these differences, we calculated a z-score for the ($\mathbf{s}_{\text{incorrect}} - \mathbf{s}_{\text{correct}}$) component of the

loss term, using the μ and σ parameters from the sampled distributions for the appropriate month. To ensure that no losses were below zero, we added the z-score of the theoretical minimum value of this component of the loss term which is -2 ($\min = \min(\mathbf{s}_{\text{incorrect}}) - \max(\mathbf{s}_{\text{correct}}) = -1 - 1 = -2$), to the difference in each month. This yielded the loss term that we optimised for the Task-Optimised agent.

A7 Calculating influence of learned variables on concept selection

While all features are scaled to fall between 0 and 1, this scaling is generally based on the theoretical minima and maxima of the feature values. Consequently, there is still variation in the regions of the parameter space which is typically occupied by concepts in a knowledge state. This means that learned parameter values and weights are not directly comparable without some transformation.

To interpret the learned feature values and weights, we take a sample of 900 calibration knowledge states (100 for each month) generated via random sampling as in the control agent. For each of these knowledge states, we obtain values of the parameters of interest and calculate the distance to the trained model's parameter values. This gives us a distribution of representative values for the distances from target variable values, that we would see when obtaining generative scores for knowledge states (i.e, the distances that would occupy vector $\mathbf{D}$ in panel 2 of Figure 2.11). We then multiply this mean distance by the learned weight for the variable, to give a representative metric for learned feature importance (i.e, how much the feature sways the probability distribution across candidate concepts).

A8 Pairwise comparison for forced choice results

	AoA probe						Control probe					
	A-M v T-O		A-M v AoA		T-O v AoA		A-M v T-O		A-M v AoA		T-O v AoA	
Month	t	p-value	t	p-value	t	p-value	t	p-value	t	p-value	t	p-value
16	4.07	$\ll 0.001^*$	4.72	$\ll 0.001^*$	0.23	0.816	8.77	$\ll 0.001^*$	9.66	$\ll 0.001^*$	0.75	0.452
17	4.54	$\ll 0.001^*$	4.27	$\ll 0.001^*$	1.16	0.247	8.63	$\ll 0.001^*$	8.01	$\ll 0.001^*$	2.05	0.041
18	5.04	$\ll 0.001^*$	3.55	$< 0.001^*$	2.55	0.012	9.71	$\ll 0.001^*$	7.73	$\ll 0.001^*$	3.39	$< 0.001^*$
19	3.14	0.002	0.31	0.755	4.54	$\ll 0.001^*$	5.15	$\ll 0.001^*$	3.01	0.003	3.01	0.003
20	2.64	0.009	1.54	0.125	4.84	$\ll 0.001^*$	5.58	$\ll 0.001^*$	3.52	$< 0.001^*$	3.10	0.002
21	3.01	0.003	0.03	0.972	3.92	$< 0.001^*$	5.62	$\ll 0.001^*$	3.39	$< 0.001^*$	3.29	0.001
22	1.87	0.062	2.53	0.012	5.11	$\ll 0.001^*$	6.12	$\ll 0.001^*$	2.82	0.005	4.48	$\ll 0.001^*$
23	0.57	0.573	4.14	$\ll 0.001^*$	3.53	$< 0.001^*$	3.59	$< 0.001^*$	0.93	0.353	3.30	0.001
24	1.44	0.152	4.25	$\ll 0.001^*$	2.47	0.014	2.70	0.001	0.07	0.947	3.25	0.001

Table A.5: Results for monthwise pairwise t-tests for forced-choice performance between each pair of model types. A-M=AoA-Matched; T-O=Task-Optimised For $\alpha = 0.05$, Bonferroni corrected threshold is $0.05/54 = 0.0009$

A9 Category analysis

To obtain the semantic categories of concepts, we started with the categorisations available in the WordBank dataset (e.g Food & Drink, Clothing, Animals), and added some additional categories to cover concepts which were not included in the AoA set but existed in the word/image embedding intersection (e.g Weapons, Medical, Tools). Upon the addition of these new categories, the early-acquired concepts were reviewed, and some items were re-assigned to these new categories if appropriate. Category distributions are given in Table A.6.

Category	Non-AoA	AoA	Total
Animals	64 (0.218)	34 (0.241)	98 (0.226)
Body parts	2 (0.007)	0 (0.000)	2 (0.005)
Clothing	19 (0.065)	17 (0.122)	36 (0.083)
Food/Drink	42 (0.143)	28 (0.199)	70 (0.161)
Furniture/Rooms	11 (0.038)	17 (0.121)	28 (0.065)
Household	51 (0.174)	19 (0.135)	70 (0.161)
Medical	6 (0.020)	0 (0.000)	6 (0.014)
Music	20 (0.068)	0 (0.000)	20 (0.046)
Outside	12 (0.041)	5 (0.035)	17 (0.039)
People	0 (0.000)	5 (0.035)	5 (0.012)
Places	5 (0.017)	1 (0.007)	6 (0.014)
Sports/Activities	24 (0.082)	0 (0.000)	24 (0.055)
Tools	6 (0.020)	2 (0.014)	8 (0.018)
Toys	1 (0.003)	4 (0.028)	5 (0.012)
Vehicles	17 (0.058)	9 (0.064)	26 (0.060)
Weapons	7 (0.024)	0 (0.000)	7 (0.016)
Other	6 (0.020)	0 (0.000)	6 (0.014)
Total	293 (1.000)	141 (1.000)	434 (1.000)

Table A.6: Frequencies of each semantic category within each concept set. Proportions in brackets represent the proportion of the concept set which is comprised of concepts from a given semantic category.

A10 Forced choice performance with AoA excluded

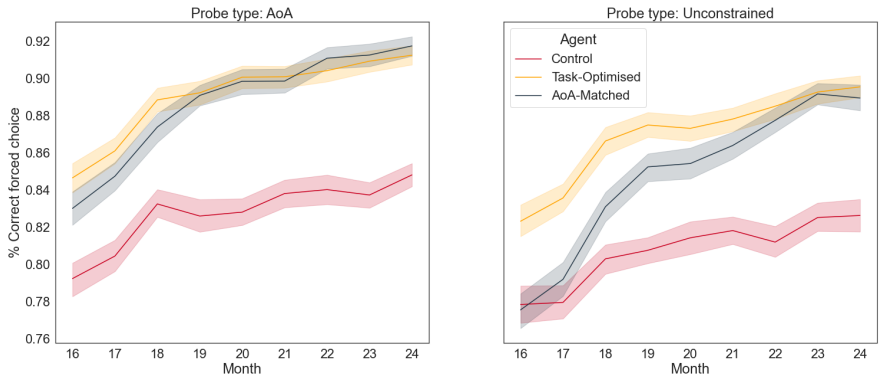


Figure A.4: Forced choice performance for control, AoA-Matched and Task-Optimised agents, when knowledge states are not permitted to contain any early-acquired concepts.

Appendix B

Supplementary information for Chapter 3

B1 Regression + Aligner model

$F(.)$ and $G(.)$ had the same structure as the MLP $F(.)$ described above for the Regression model. Further to the description in the main text, cycle loss $\mathcal{L}_{cyc}$ included the parallel loss term for the mapping of all $\boldsymbol{Y}$ to $\boldsymbol{Y}''$. This makes the total cycle consistency loss:

$$\bar{\mathcal{L}}_{cyc} = \frac{1}{2}\left(\mathbb{E}_x\left[\|\boldsymbol{X} - \boldsymbol{X}''\|\right] + \mathbb{E}_y\left[\|\boldsymbol{Y} - \boldsymbol{Y}''\|\right]\right)$$

As described in the main text, $\mathcal{L}_{dist}$ is the mean negative log likelihood (NLL) of $F(\boldsymbol{X})$ as samples from a Gaussian mixture model comprised of 2D Gaussian kernels placed on Y (GMM_Y). The Gaussian mixture model is defined as follows:

$$GMM_Y = \frac{1}{6}\sum_{j=1}^{6}\mathcal{N}(\mathbf{y}; \mathbf{y}_j, \sigma\boldsymbol{I}_2),$$

where σ=0.1. The total distribution loss includes the mean NLL of $\boldsymbol{Y}''$ as a sample from GMM_Y and the mean NLL of $\boldsymbol{X}''$ as a sample from GMM_X.

As both $\mathcal{L}_{cyc}$ and $\mathcal{L}_{dist}$ required exposure to the whole space of stimuli, the unsupervised loss terms were not introduced until after the first block of passive

trials in model training, where $t > 6$. λ_{cyc} and λ_{dist} specified the weights of the cycle consistency and distribution loss terms respectively, relative to the supervised loss term. On each trial in model training, the total loss term was:

$$\mathcal{L} = \frac{1}{2}\left(NLL_{\boldsymbol{y}_t}(\boldsymbol{x}'_t) + NLL_{\boldsymbol{x}_t}(\boldsymbol{y}'_t)\right) + \lambda_{cyc}\bar{\mathcal{L}}_{cyc} + \lambda_{dist}\bar{\mathcal{L}}_{dist}$$

where :

$$\lambda_{cyc} = \begin{cases} \lambda_{cyc}, & \text{if } t > 6 \\ 0, & \text{otherwise} \end{cases} \qquad \lambda_{dist} = \begin{cases} \lambda_{dist}, & \text{if } t > 6 \\ 0, & \text{otherwise} \end{cases}$$

Appendix C

Supplementary information for Chapter 4

C1 Algorithm details

C1.1 Monte Carlo Tree Search

Each iteration of the MCTS algorithm begins from the root node. At each node, a child node is selected based on some function of the child node utility, until either (i) a node which has not been fully expanded or (ii) a terminal node is reached. Upon selecting an unexpanded node, the algorithm enters a simulation stage, sometimes referred to as *rollout* or *playout*. Here, the value of the selected nodes are estimated by simulating n_s further paths through the tree.

Once each simulation phase has been completed, here meaning that each item has been mapped to an item in the other space, the value of the terminal state is calculated. The highest mapping value across the n_s simulations is backpropagated via the nodes visited prior to expansion, and is used to update their estimates.

When the algorithm is terminated - either by interruption or by computational budget being reached - actions within the output tree can be selected by a range of mechanisms (Browne et al., 2012; Schadd et al., 2008; Chaslot et al.,

2008). We use the *max-child* policy, choosing the action with the maximum estimated reward.

The *tree policy* refers to the decision policy used to select a node at each level of the tree where node is unexplored. Kocsis and Szepesvári (2006) proposes a node tree policy based on the UCB1 score. This score balances the need to explore nodes which have not been frequently visited with the exploitation of current value estimates. The UCB1 score for node j is shown in equation C1.1:

$$\text{UCB1} = \frac{\sum v_j}{n_j} + c\sqrt{\frac{\ln(n)}{n_j}}, \tag{C1.1}$$

where n_j is the number of times node j has been visited from the parent node in question, and n is the total number of times the parent node of j has been visited. $\sum v_j$ is the sum of all values obtained when passing through node j on previous visits. The constant c in the UCB1 formula controls the balance of exploration and exploitation of existing knowledge of high-value states.

An adapted version of the UCB1 score, which was proposed in Schadd et al. (2008), can be used for single-player games where the game outcome does not lie within a preset interval and cannot be summarised as a draw, win or loss. This modified score is shown in equation C1.2, and is the score used in our MCTS procedure.

$$\text{UCB1} = \frac{\sum v_j}{n_j} + c\sqrt{\frac{\ln(n)}{n_j}} + \sqrt{\frac{\sum v_j^2 - \left(\frac{\sum v_j}{n_j}\right)^2 n_j + D}{n_j}}, \tag{C1.2}$$

Where D is a high constant to ensure that nodes which have rarely been expxlored are viewed as uncertain.

Once a node is selected for expansion, simulated play continues until a terminal node is reached. For our purposes, a terminal node is one in which all items in space X are mapped onto items in space Y. In the simulation phase, actions are selected according to a separate decision policy, termed the *default policy* (Browne et al., 2012). The default policy may be random - where the

probability of selection is an even distribution across actions - or based on some prior estimate of node values.

C1.2 Kuhn-Munkres algorithm

The process of the Kuhn-Munkres algorithm, in matrix formulation, is provided below:

- **Step 1:** Subtract each row's smallest value from all row items, such that each row's minimum cost assignment takes the value 0. If resultant matrix can be used for assigning (i.e., if there is one 0 per row and per column), terminate.
- **Step 2:** Else, repeat step 1 column-wise (subtract each column's smallest value from all column items). If resultant matrix can be used for assigning (i.e., if there is one 0 per row and per column), terminate.
- **Step 3:** Else, try to randomly 'assign' one zero in each row by starring it. Zeros cannot be assigned if they are in a column where an assigned zero already exists.
- **Step 4:** Obscure all columns which contain an assigned (or starred) zero. Then, Find an unobscured zero and prime it. If all zeros are obscured, skip to Step 5.
  - If the primed zero has a starred zero on the same row, unobscure the column containing the starred zero and obscure the whole row. Then, return to beginning of step 4.
  - Else, if the primed zero has no starred zero on the same row:
    - Take a path from this zero as follows: (a) Find a starred zero in the corresponding column. If one exists, proceed to (b). Else, stop. (b) Find a primed zero in the corresponding row. This will always exist. Then, return to (a)

    - For all zeros encountered while making this path, star any primed zero and unstar any starred zeros.
    - Unprime all primed zeros, and unobscure all lines
    - Continue looping step 4 until conditions above permit skipping to step 5

- **Step 5:** Zeros are now obscured by the minimum number of lines. Now, find the lowest uncovered value in the matrix. Subtract this from every unmarked matrix element, and add it to every element obscured by a row line and a column line (effectively, this subtracts the number from all unobscured rows and adds the same number to all obscured columns). Repeat steps 4 and 5 until the minimum number of lines equals min(rows, columns). The resultant starred zeros indicate the optimal assignment.

C1.3 Exhaustive start implementation

The process for implementing the exhaustive start selection policy is as follows:

- Start by exhaustively searching the set of mappings for the first k concepts
- The number of sets to search is $\binom{N}{k}$, where N is the total number of concepts and k is the size of the mapping
- This is quickly very computationally demanding. E.g., for 50 concepts, to search the space of the first 3 concepts requires running simulations for 19,600 mappings of the first 3 concepts from system X. In these simulations, we map the first 3 concepts.
- Each mapping has to be tested through multiple simulations (here, we use 20 simulations), and after mapping those 3 concepts we then have to continue performing MCTS to a sufficient level for the results to be meaningful.

C2 Alignment algorithm testing

C2.1 Self-self mapping

Predictor	df	F	p	η_p^2
N_u	1	3.97	0.048*	0.03
Algorithm	6	40.92	< 0.001*	0.65
N_u x Algorithm	6	3.26	0.005*	0.13

Table C.1: Results of the 2-way ANOVA for the experiment on unsupervised self-self mapping performance of algorithms. N_u is the number of 'unseen' concepts, i.e. the number of concepts for which the algorithm learns the mapping. Significance at level $\alpha = 0.05$ is indicated by *.

N concepts	Algorithm	t	df	p
10	mctsbasicRandom	3.473	9	0.007
	mctsheuristicRandom	2.806	9	0.021
	mctsbasicConstrained	3.601	9	0.006
	mctsheuristicConstrained	4.243	9	0.002*
	mctsexhaustiveConstrained	4.830	9	0.001*
	cycle	1.641	9	0.132
	kuhn	∞	9	< 0.001*
50	mctsbasicRandom	0.709	9	0.496
	mctsheuristicRandom	0.739	9	0.479
	mctsbasicConstrained	0.000	9	1.000
	mctsheuristicConstrained	-0.612	9	0.555
	mctsexhaustiveConstrained	4.813	9	0.001*
	cycle	1.748	9	0.111
	kuhn	∞	9	< 0.001*

Table C.2: Table of t-test results for difference from chance for each algorithm in the task of unsupervised self-self mapping. Bonferroni-corrected significance level = 0.05/14 = 0.004. Significance at this level is indicated by *.

C2.2 Self-self mapping with noise

Predictor	df	F	p	η_p^2
N_u	1	13.12	< .001*	0.03
Algorithm	6	103	< .001*	0.57
noise	1	0.25	0.615	0 .00
N_u x Algorithm	6	3.76	< .001*	0.05
N_u x noise	1	0.23	0.629	0.00
Algorithm x noise	6	34.51	< .001*	0.31
N_u x noise x Algorithm	6	0.24	0.964	0.00

Table C.3: ANOVA table for the results of the experiment on unsupervised performance of algorithms for representations for noisy versions of themself. N_u is the number of 'unseen' concepts, i.e. the number of concepts for which the algorithm learns the mapping. Noise size indicates the amount of noise added to the representations for which the mappings were learned. Significance at level $\alpha = 0.05$ is indicated by *.

	Noise size =	0.01			0.03			0.1		
N concepts	Algorithm	t	df	p	t	df	p	t	df	p
10	cycle	1.166	10	0.271	2.324	10	0.042	1.000	10	0.341
	kuhn	35.404	10	$< .001^*$	9.461	10	$< .001^*$	2.206	10	0.052
	mctsbasicConstrained	1.857	9	0.096	2.535	9	0.032	2.167	9	0.058
	mctsbasicRandom	1.868	9	0.095	3.143	9	0.012	1.078	9	0.309
	mctsexhaustiveConstrained	8.232	9	$< .001^*$	10.091	9	$< .001^*$	10.474	9	$< .001^*$
	mctsheuristicConstrained	1.922	9	0.087	0.802	9	0.443	0.000	9	1.000
	mctsheuristicRandom	3.508	19	$< .001^*$	4.925	19	$< .001^*$	2.557	19	0.019
50	cycle	-1.399	10	0.192	0.000	10	1.000	1.000	10	0.341
	kuhn	33.001	10	$< .001^*$	20.499	10	$< .001^*$	7.321	10	$< .001^*$
	mctsbasicConstrained	0.199	10	0.846	-0.429	9	0.678	-1.000	9	0.343
	mctsbasicRandom	-0.802	9	0.443	0.669	9	0.520	0.514	9	0.619
	mctsexhaustiveConstrained	7.216	9	$< .001^*$	8.060	9	$< .001^*$	7.060	9	$< .001^*$
	mctsheuristicConstrained	-0.614	10	0.553	0.000	9	1.000	1.500	9	0.168
	mctsheuristicRandom	-1.453	19	0.163	0.000	19	1.000	-1.000	19	0.330

Table C.4: Table of t-test results for difference from chance for each algorithm in the task of unsupervised mapping of a systems to a noisy version of itself. N_u is the number of 'unseen' concepts, i.e. the number of concepts for which the algorithm learns the mapping. Noise size indicates the amount of noise added to the representations for which the mappings were learned. Bonferroni-corrected significance level = 0.05/42 = 0.001. Significance at this level is indicated by *.

C2.3 Unsupervised visual-linguistic mapping

Predictor	df	F	p	η_p^2
N_u	1	4.24	0.040^*	0.01
setting_type	2	0.17	0.844	0.00
Algorithm	6	32.54	$< 0.001^*$	0.34
N_u x setting	2	0.01	0.991	0.00
N_u x Algorithm	6	2.56	0.019	0.04
setting_type x Algorithm	12	1.34	0.194	0.04
N_u x setting x Algorithm	12	1.76	0.054	0.05

Table C.5: ANOVA table for the results of the experiment on unsupervised performance of algorithms for mapping between image and word embeddings. Greenhouse-Geisser correction has been applied to corresct for violations of Sphericity assumptions. Setting was a factor specifying which alignment scoring setting was used for the mapping algorithm. Significance at level $\alpha = 0.05$ is indicated by *.

	setting	bestCorr			bestProp			Original settings		
N concepts	Algorithm	t	df	p	t	df	p	t	df	p
10	cycle	0.803	10	0.441	1.936	10	0.082	1.456	10	0.176
	kuhn	-0.559	10	0.588	0.000	10	1.000	1.102	10	0.296
	mctsbasicConstrained	0.557	9	0.591	1.784	9	0.108	0.000	9	1.000
	mctsbasicRandom	0.840	9	0.423	-1.627	9	0.138	-1.152	9	0.279
	mctsexhaustiveConstrained	26.509	54	$< 0.001^*$	27.507	54	$< 0.001^*$	11.619	9	$< 0.001^*$
	mctsheuristicConstrained	0.318	9	0.758	0.605	9	0.560	1.561	9	0.153
	mctsheuristicRandom	2.333	9	0.045	0.709	9	0.496	0.000	9	1.000
50	cycle	0.000	10	1.000	0.000	10	1.000	2.206	10	0.052
	kuhn	-1.174	10	0.267	1.936	10	0.082	0.614	10	0.553
	mctsbasicConstrained	-1.809	9	0.104	-0.896	9	0.394	-0.314	7	0.763
	mctsbasicRandom	-0.802	9	0.443	-1.861	9	0.096	0.429	9	0.678
	mctsexhaustiveConstrained	16.474	54	$< 0.001^*$	11.776	54	$< 0.001^*$	19.744	54	$< 0.001^*$
	mctsheuristicConstrained	1.809	9	0.104	-2.236	9	0.052	0.000	9	1.000
	mctsheuristicRandom	0.000	9	1.000	0.000	9	1.000	-0.557	9	0.591

Table C.6: Table of t-test results for difference from chance for each algorithm in the task of unsupervised mapping between image and word embeddings. N_u is the number of 'unseen' concepts, i.e. the number of concepts for which the algorithm learns the mapping. Noise size indicates the amount of noise added to the representations for which the mappings were learned. Bonferroni-corrected significance level = 0.05/42 = 0.001. Significance at this level is indicated by *.

C2.4 Supervised visual-linguistic mapping

Predictor	df	F	p	η_p^2
N_u	1	5.27	0.022*	0.01
Algorithm	6	1.44	0.198	0.02
N_s	2	19.8	0.001*	0.09
N_u x Algorithm	6	0.45	0.847	0.01
N_u x N_s	2	4.83	0.008*	0.02
Algorithm x N_s	12	4.81	0.001*	0.13
N_u x Algorithm x N_s	12	1.54	0.108	0.05

Table C.7: ANOVA table for the results of the experiment on supervised performance of algorithms for mapping visual- to linguistic representations. Greenhouse-Geisser correction has been applied to corresct for violations of Sphericity assumptions. N_s is the number of 'seen' concepts, i.e. the number of concepts for which supervision is provided. N_u is the number of 'unseen' concepts, i.e. the number of concepts for which the algorithm learns the mapping. Significance at level $\alpha = 0.05$ is indicated by *.

	N_sup=	5			10			50		
N concepts	Algorithm	t	df	p	t	df	p	t	df	p
10	cycle	1.861	9	0.096	3.515	9	0.007	7.141	9	$< 0.001^*$
	kuhn	1.809	9	0.104	4.993	9	0.001^*	5.526	9	$< 0.001^*$
	mctsbasicConstrained	1.246	9	0.244	1.309	9	0.223	0.429	9	0.678
	mctsbasicRandom	0.612	9	0.555	1.168	9	0.273	1.069	8	0.316
	mctsheuristicConstrained	-0.557	9	0.591	-0.208	9	0.840	2.946	9	0.016
	mctsheuristicRandom	0.246	9	0.811	0.361	9	0.726	-0.318	9	0.758
	regression	1.784	9	0.108	3.354	9	0.008	1.778	9	0.109
50	cycle	2.167	9	0.058	3.515	9	0.007	6.520	9	$< 0.001^*$
	kuhn	4.707	9	0.001^*	5.277	9	0.001^*	10.115	9	$< 0.001^*$
	mctsbasicConstrained	0.688	9	0.509	-0.896	9	0.394	0.896	9	0.394
	mctsbasicRandom	-0.688	9	0.509	-0.287	9	0.780	-1.500	9	0.168
	mctsheuristicConstrained	1.177	9	0.269	-0.688	9	0.509	0.480	9	0.642
	mctsheuristicRandom	0.429	9	0.678	-1.000	9	0.343	-0.688	9	0.509
	regression	2.167	9	0.058	2.012	9	0.075	3.584	9	0.006

Table C.8: Table of t-test results for difference from chance for each algorithm in the task of supervised visual- to linguistic mapping. Bonferroni-corrected significance level = 0.05/42 = 0.001. Significance at this level is indicated by *.

C3 Alignment prior results

C3.1 ANOVA results for zero-shot performance

Predictor	df	F	p	η^2
Supervision	$(3, 00, 75.00)$	0.109	0.954	0.002
Prior	$(1.64, 122.64)$	1227.288	$< .001^*$	0.876
Supervision * Prior	$(4.91, 122.64)$	0.973	0.436	0.016

Table C.9: Results of a 2-way mixed ANOVA for the uplift in zero-shot classification performance over a classifier with a uniform prior. Greenhouse-Geisser correction has been applied to coresct for violations of Sphericity assumptions. * Indicates a significant result for $\alpha = 0.05$.

C3.2 Post-hoc pairwise comparisons for zero-shot performance

	Prior_i	Prior_j	t	df	p_{adj}
25%	Cycle + Reg	Reg only	3.25	19	0.013*
	Cycle + Reg	Similarity corr	27.33	19	< 0.001*
	Reg only	Similarity corr	25.93	19	< 0.001*
50%	Cycle + Reg	Reg only	2.07	19	0.156
	Cycle + Reg	Similarity corr	25.70	19	< 0.001*
	Reg only	Similarity corr	23.18	19	< 0.001*
75%	Cycle + Reg	Reg only	4.31	19	0.001*
	Cycle + Reg	Similarity corr	25.68	19	< 0.001*
	Reg only	Similarity corr	24.37	19	< 0.001*
100%	Cycle + Reg	Reg only	-0.84	19	1.000
	Cycle + Reg	Similarity corr	26.01	19	< 0.001*
	Reg only	Similarity corr	11.38	19	< 0.001*

Table C.10: Post-hoc pairwise comparisons for uplift in accuracy relative to uniform prior in 100-way zero-shot classification. * Indicates a significant result for $\alpha = 0.05$

C3.3 ANOVA results for few-shot performance

Predictor	df	F	p	η^2
Supervision	$(3, 00, 73.00)$	1.363	0.261	0.027
Prior	$(1.65, 120.29)$	46.901	< .001*	0.103
N-Shot	$(1.83, 133.55)$	13.896	< .001*	0.043
Supervision * Prior	$(4.94, 120.29)$	1.318	0.261	0.010
Supervision * N-shot	$(5.49, 133.55)$	1.096	0.367	0.010
Prior * N-shot	$(3.27, 238.70)$	2.248	0.078	0.003
Supervision * Prior * N-shot	$(9.81, 238.70)$	0.986	0.455	0.004

Table C.11: Results of a 3-way mixed ANOVA for the uplift in classification performance over a classifier with a uniform prior. Greenhouse-Geisser correction has been applied to corresct for violations of Sphericity assumptions. * Indicates a significant result for $\alpha = 0.05$.

C3.4 Post-hoc pairwise comparisons for few-shot performance

Shot	Supervision %	Prior_i	Prior_j	t	df	p_{adj}
1	25	Cycle + reg	Reg only	0.465	19	1.000
		Cycle + reg	Similarity corr	3.885	19	0.003*
		Reg only	Similarity corr	2.800	19	0.034*
	50	Cycle + reg	Reg only	2.230	19	0.114
		Cycle + reg	Similarity corr	3.049	19	0.02*
		Reg only	Similarity corr	1.313	19	0.615
	75	Cycle + reg	Reg only	-0.011	19	1.000
		Cycle + reg	Similarity corr	3.622	19	0.005*
		Reg only	Similarity corr	3.666	19	0.005*
	100	Cycle + reg	Reg only	-1.329	18	0.603
		Cycle + reg	Similarity corr	1.658	18	0.345
		Reg only	Similarity corr	2.446	18	0.075
2	25	Cycle + reg	Reg only	0.424	18	1.000
		Cycle + reg	Similarity corr	5.232	18	0.000
		Reg only	Similarity corr	4.270	18	0.001*
	50	Cycle + reg	Reg only	2.063	18	0.161
		Cycle + reg	Similarity corr	2.842	18	0.032*
		Reg only	Similarity corr	2.152	18	0.136
	75	Cycle + reg	Reg only	-0.517	18	1.000
		Cycle + reg	Similarity corr	2.828	18	0.033*
		Reg only	Similarity corr	5.410	18	0.000*
	100	Cycle + reg	Reg only	0.624	17	1.000
		Cycle + reg	Similarity corr	2.788	17	0.038*
		Reg only	Similarity corr	2.699	17	0.046*
5	25	Cycle + reg	Reg only	0.013	18	1.000
		Cycle + reg	Similarity corr	10.761	18	0.000*
		Reg only	Similarity corr	6.986	18	0.000*
	50	Cycle + reg	Reg only	1.109	18	0.846
		Cycle + reg	Similarity corr	2.404	18	0.082
		Reg only	Similarity corr	2.682	18	0.046*
	75	Cycle + reg	Reg only	-0.365	18	1.000
		Cycle + reg	Similarity corr	6.331	18	0.000*
		Reg only	Similarity corr	6.335	18	0.000*
	100	Cycle + reg	Reg only	0.481	17	1.000
		Cycle + reg	Similarity corr	5.160	17	0.000*
		Reg only	Similarity corr	5.004	17	0.000*

Table C.12: Post-hoc pairwise comparisons for uplift in accuracy relative to uniform prior in 100-way 1-, 2- and 5-shot classification. * Indicates a significant result for $\alpha = 0.05$

Bibliography

C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Database TheoryICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8*, pages 420–434. Springer, 2001.

K. Aho, B. D. Roads, and B. C. Love. System alignment supports cross-domain learning and zero-shot generalisation. *Cognition*, 227:105200, 2022.

Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2927–2936, 2015.

N. Akhtar. The robustness of learning through overhearing. *Developmental Science*, 8(2):199–209, 2005.

N. Akhtar, J. Jipson, and M. A. Callanan. Learning words through overhearing. *Child development*, 72(2):416–430, 2001.

E. M. Aminoff, K. Kveraga, and M. Bar. The role of the parahippocampal cortex in cognition. *Trends in cognitive sciences*, 17(8):379–390, 2013.

M. Andrews, G. Vigliocco, and D. Vinson. Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116 (3):463, 2009.

M. Antoniak and D. Mimno. Evaluating the stability of embedding-based word

similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119, 2018.

M. Artetxe, G. Labaka, and E. Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2289–2294, 2016.

R. K. Bakshi, N. Kaur, R. Kaur, and G. Kaur. Opinion mining and sentiment analysis. In *2016 3rd international conference on computing for sustainable global development (INDIACom)*, pages 452–455. IEEE, 2016.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

X. Bao, E. Gjorgieva, L. K. Shanahan, J. D. Howard, T. Kahnt, and J. A. Gottfried. Grid-like neural representations support olfactory navigation of a two-dimensional odor space. *Neuron*, 102(5):1066–1075, 2019.

M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.

D. N. Barry and B. C. Love. A neural network account of memory replay and knowledge consolidation. *Cerebral Cortex*, 33(1):83–95, 2023.

L. W. Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645, 2008.

L. W. Barsalou, W. K. Simmons, A. K. Barbey, and C. D. Wilson. Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2):84–91, 2003.

F. C. Bartlett and F. C. Bartlett. *Remembering: A study in experimental and social psychology*. Cambridge university press, 1995.

T. E. Behrens, T. H. Muller, J. C. Whittington, S. Mark, A. B. Baram, K. L. Stachenfeld, and Z. Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018.

R. Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

J. L. Bellmund, P. Gärdenfors, E. I. Moser, and C. F. Doeller. Navigating cognition: Spatial codes for human thinking. *Science*, 362(6415):eaat6766, 2018.

E. M. Bender. Thought experiment in the national library of thailand, 2023. URL `https://medium.com/@emilymenonbender/thought-experiment-in-the-national-library-of-thailand-f2bf761a8a83`. Accessed: 05/07/2023.

E. M. Bender and A. Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, 2020.

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

L. Bloom. One word at a time. In *One Word at a Time*. De Gruyter Mouton, 2013.

M. F. Bonner and R. A. Epstein. Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, 12(1):1–16, 2021.

M. Braginsky, D. Yurovsky, V. A. Marchman, and M. Frank. From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. In *CogSci*, 2016.

C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton. A survey of

monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.

E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, 2012.

E. Bruni, N.-K. Tran, and M. Baroni. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47, 2014.

A. Caramazza, A. E. Hillis, B. C. Rapp, and C. Romani. The multiple semantics hypothesis: Multiple confusions? *Cognitive neuropsychology*, 7(3): 161–189, 1990.

S. Carey and E. Bartlett. Acquiring a single new word. 1978.

E. A. Cartmill, B. F. Armstrong, L. R. Gleitman, S. Goldin-Meadow, T. N. Medina, and J. C. Trueswell. Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, 110(28):11278–11283, 2013.

G. M. J. Chaslot, M. H. Winands, H. J. V. D. HERIK, J. W. Uiterwijk, and B. Bouzy. Progressive strategies for monte-carlo tree search. *New Mathematics and Natural Computation*, 4(03):343–357, 2008.

C.-h. Chen and C. Yu. Grounding statistical learning in context: The effects of learning and retrieval contexts on cross-situational word learning. *Psychonomic bulletin & review*, 24(3):920–926, 2017.

L. Chen, Z. Gan, Y. Cheng, L. Li, L. Carin, and J. Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR, 2020a.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020b.

T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020c.

Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020d.

E. M. Clerkin and L. B. Smith. Real-world statistics at two timescales and a mechanism for infant learning of object names. *Proceedings of the National Academy of Sciences*, 119(18):e2123239119, 2022.

E. M. Clerkin, E. Hart, J. M. Rehg, C. Yu, and L. B. Smith. Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160055, 2017.

A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

A. O. Constantinescu, J. X. OReilly, and T. E. Behrens. Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468, 2016.

B. Dai, S. Fidler, R. Urtasun, and D. Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE international conference on computer vision*, pages 2970–2979, 2017.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

C. Demircan, T. Saanum, L. Pettini, M. Binz, B. M. Baczkowski, P. Kaanders, C. F. Doeller, M. M. Garvert, and E. Schulz. Language aligned visual representations predict human behavior in naturalistic learning tasks. *arXiv preprint arXiv:2306.09377*, 2023.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

L. A. Doumas, G. Puebla, A. E. Martin, and J. E. Hummel. Relation learning in a neurocomputational architecture supports cross-domain transfer. *arXiv preprint arXiv:1910.05065*, 2019.

S. Edelman. Representation is representation of similarities. *Behavioral and brain sciences*, 21(4):449–467, 1998.

T. Eerola and J. K. Vuoskoski. A review of music and emotion studies: Approaches, emotion models, and stimuli. *Music Perception: An Interdisciplinary Journal*, 30(3):307–340, 2012.

J. L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

N. Enfield. *Language vs. Reality: Why Language Is Good for Lawyers and Bad for Scientists*. MIT Press, 2022. ISBN 9780262046619. URL `https://books.google.co.uk/books?id=DtOMEAAAQBAJ`.

H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

I. Farkaš. Indispensability of computational modeling in cognitive science. *Journal of Cognitive Science*, 13(4):401–429, 2012.

Y. Feng, L. Ma, W. Liu, and J. Luo. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134, 2019.

L. Fenson, P. S. Dale, J. S. Reznick, E. Bates, D. J. Thal, S. J. Pethick, M. Tomasello, C. B. Mervis, and J. Stiles. Variability in early communicative

development. *Monographs of the society for research in child development*, pages i–185, 1994.

L. Fenson et al. *MacArthur-Bates communicative development inventories*. Paul H. Brookes Publishing Company Baltimore, MD, 2007.

L. Fernandino, J. R. Binder, R. H. Desai, S. L. Pendl, C. J. Humphries, W. L. Gross, L. L. Conant, and M. S. Seidenberg. Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral cortex*, 26(5):2018–2034, 2016.

J. R. Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

A. Fourtassi and E. Dupoux. The role of word-word co-occurrence in word learning. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, pages 662–667, 2016.

M. C. Frank, M. Braginsky, D. Yurovsky, and V. A. Marchman. Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, 44(3):677, 2017.

A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. 2013.

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

A. Gampe, K. Liebal, and M. Tomasello. Eighteen-month-olds learn novel words through overhearing. *First Language*, 32(3):385–397, 2012.

D. Ganguly, D. Roy, M. Mitra, and G. J. Jones. Word embedding based generalized language model for information retrieval. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 795–798, 2015.

D. Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983.

D. Gentner and K. J. Holyoak. Reasoning and learning by analogy: Introduction. *American psychologist*, 52(1):32, 1997.

D. Gentner and L. Smith. Analogical reasoning. *Encyclopedia of human behavior*, 2:130–136, 2012.

M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214–224, 2017.

B. A. Goldfield and J. S. Reznick. Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of child language*, 17(1):171–183, 1990.

R. L. Goldstone and D. L. Medin. Time course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(1):29, 1994.

R. L. Goldstone and B. J. Rogosky. Using relations within conceptual systems to translate across conceptual systems. *Cognition*, 84(3):295–320, 2002.

U. Goswami. Children's use of analogy in learning to read: A developmental study. *Journal of experimental child psychology*, 42(1):73–83, 1986.

A. Greve, E. Cooper, R. Tibon, and R. N. Henson. Knowledge is power: Prior knowledge aids memory for both congruent and incongruent events, but in different ways. *Journal of Experimental Psychology: General*, 148(2):325, 2019.

J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang. Unpaired image captioning via scene graph alignments. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10322–10331, 2019. doi: 10.1109/ICCV.2019.01042.

A. Hagberg, P. Swart, and D. S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, volume 2016, page 2116. NIH Public Access, 2016.

S. Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

T. H. Heibeck and E. M. Markman. Word learning in children: An examination of fast mapping. *Child development*, pages 1021–1034, 1987.

T. Hills. The company that words keep: comparing the statistical structure of child-versus adult-directed language. *Journal of child language*, 40(3): 586–604, 2013.

T. T. Hills, M. Maouene, J. Maouene, A. Sheya, and L. Smith. Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological science*, 20(6):729–739, 2009.

J. R. Hodges, N. Graham, and K. Patterson. Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory*, 3(3-4):463–495, 1995.

K. J. Holyoak. Analogy and relational reasoning. 2012.

K. J. Holyoak and P. Thagard. Analogical mapping by constraint satisfaction. *Cognitive science*, 13(3):295–355, 1989.

H. Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992.

Q. Huang and H. Luo. Shared structure facilitates working memory of multiple sequences via neural replay. *bioRxiv*, pages 2023–07, 2023.

S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, Q. Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

V. K. Jaswal and E. M. Markman. Learning proper and common names in inferential versus ostensive contexts. *Child Development*, 72(3):768–786, 2001.

E. Jefferies. The neural basis of semantic cognition: converging evidence from neuropsychology, neuroimaging and tms. *Cortex*, 49(3):611–625, 2013.

B. T. Johns and M. N. Jones. Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, 4(1):103–120, 2012.

P. N. Juslin, J. A. Sloboda, et al. Music and emotion. *D. DEUTSCH (Org.)*, 2001.

A. Kappes and T. Sharot. The automatic nature of motivated belief updating. *Behavioural Public Policy*, 3(1):87–103, 2019.

H. Karmazyn-Raz and L. B. Smith. Discourse with few words: coherence statistics, parent-infant actions on objects, and object names. *Language Acquisition*, pages 1–19, 2022.

M. Kiefer and F. Pulvermüller. Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *cortex*, 48 (7):805–825, 2012.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

L. Kocsis and C. Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.

W. Köhler. *Gestalt psychology: An introduction to new concepts in modern psychology*, volume 18. WW Norton & Company, 1970.

R. G. Kuehni. How many object colors can we distinguish? *Color Research & Application*, 41(5):439–444, 2016.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

M.-A. Lachaux, B. Roziere, L. Chanussot, and G. Lample. Unsupervised translation of programming languages. *arXiv preprint arXiv:2006.03511*, 2020.

I. Laina, C. Rupprecht, and N. Navab. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7414–7424, 2019.

B. M. Lake and G. L. Murphy. Word meaning in minds and machines. *Psychological Review*, 2021.

B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

M. A. Lambon Ralph. Neurocognitive insights on conceptual knowledge and its breakdown. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1634):20120392, 2014.

T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

M. E. Lassaline and G. L. Murphy. Alignment and category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(1):144, 1998.

M. A. Lawrence and M. M. A. Lawrence. Package ez. *R package version*, 4 (0), 2016.

A. Lazaridou, E. Bruni, and M. Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, 2014.

A. Lazaridou, G. Chrupała, R. Fernández, and M. Baroni. Multimodal semantic learning from child-directed input. In *Knight K, Nenkova A, Rambow O, editors. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016 Jun 12-17; San Diegio, California. Stroudsburg (PA): Association for Computational Linguistics; 2016. p. 387–92.* ACL (Association for Computational Linguistics), 2016.

Y. C. Leong, B. L. Hughes, Y. Wang, and J. Zaki. Neurocomputational mechanisms underlying motivated seeing. *Nature human behaviour*, 3(9):962–973, 2019.

M. Lewis, M. Zettersten, and G. Lupyan. Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, 116 (39):19237–19238, 2019.

P. Li, C. Burgess, and K. Lund. The acquisition of word meaning through global lexical co-occurrences. In *Proceedings of the thirtieth annual child language research forum*, pages 166–178. Citeseer, 2000.

E. V. Lieven. Crosslinguistic and crosscultural aspects of language addressed to children. 1994.

G. W. Lindsay. Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience*, 14:29, 2020.

Q. Liu and G. Lupyan. Cross-domain semantic alignment: concrete concepts are more abstract than you think. *Philosophical Transactions of the Royal Society B*, 378(1870):20210372, 2023.

M. M. Louwerse. Embodied relations are encoded in language. *Psychonomic Bulletin & Review*, 15:838–844, 2008.

M. M. Louwerse. Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in Cognitive Science*, 10(3):573–589, 2018.

H. Lu, D. Chen, and K. J. Holyoak. Bayesian analogy with relational transformations. *Psychological review*, 119(3):617, 2012.

H. Lu, Y. N. Wu, and K. J. Holyoak. Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, 116(10):4176–4181, 2019a.

J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019b.

K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203–208, 1996.

M. R. Luo and B. Rigg. Chromaticity-discrimination ellipses for surface colours. *Color Research & Application*, 11(1):25–42, 1986.

M. R. Luo, G. Cui, and C. Li. Uniform colour spaces based on CIECAM02 colour appearance model. *Color Research & Application*, 31(4):320–330, 2006.

B. MacWhinney. The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database, 2000.

R. Marjieh, P. Van Rijn, I. Sucholutsky, T. Sumers, H. Lee, T. L. Griffiths, and N. Jacoby. Words are all you need? language as an approximation for human similarity judgments. In *The Eleventh International Conference on Learning Representations*, 2022.

R. Marjieh, N. Jacoby, J. C. Peterson, and T. L. Griffiths. The universal law of generalization holds for naturalistic stimuli. *arXiv preprint arXiv:2306.08564*, 2023.

A. B. Markman and D. Gentner. The effects of alignability on memory. *Psychological Science*, pages 363–367, 1997.

E. M. Markman. Constraints children place on word meanings. *Cognitive science*, 14(1):57–77, 1990.

E. M. Markman. Constraints on word meaning in early language acquisition. *Lingua*, 92:199–227, 1994.

A. Martin. The representation of object concepts in the brain. *Annu. Rev. Psychol.*, 58:25–45, 2007.

A. Martin. Grapesgrounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic bulletin & review*, 23(4):979–990, 2016.

C. B. Martin, D. Douglas, R. N. Newsome, L. L. Man, and M. D. Barense. Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *Elife*, 7:e31873, 2018.

J. Mayor and K. Plunkett. A statistical estimate of infant and toddler vocabulary size from cdi analysis. *Developmental Science*, 14(4):769–785, 2011.

D. McCarthy. Language development in children. In *Manual of child psychology.*, pages 476–581. John Wiley & Sons Inc, 1946.

J. L. McClelland. The place of modeling in cognitive science. *Topics in Cognitive Science*, 1(1):11–38, 2009.

K. McCormick, J. Kim, S. M. List, and L. C. Nygaard. Sound to meaning mappings in the bouba-kiki effect. In *CogSci*, volume 2015, pages 1565–1570, 2015.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

K. McRae, V. R. De Sa, and M. S. Seidenberg. On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126(2):99, 1997.

M. Melgosa, E. Hita, A. Poza, D. H. Alman, and R. S. Berns. Suprathreshold color-difference ellipsoids for surface colors. *Color Research & Application*, 22(3):148–155, 1997.

L. Meteyard, S. R. Cuadrado, B. Bahrami, and G. Vigliocco. Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, 48(7): 788–804, 2012.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013b.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013c.

M. Mitchell. Abstraction and analogy-making in artificial intelligence. *arXiv preprint arXiv:2102.10717*, 2021.

T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

A. Mogadala, M. Kalimuthu, and D. Klakow. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *Journal of Artificial Intelligence Research*, 71:1183–1317, 2021.

R. M. Mok and B. C. Love. A non-spatial account of place and grid cells based on clustering models of concept learning. *Nature communications*, 10 (1):1–9, 2019.

W. Mokrzycki and M. Tatol. Colour difference delta E - a survey. *Mach. Graph. Vis*, 20(4):383–411, 2011.

N. Moroney, M. D. Fairchild, R. W. Hunt, C. Li, M. R. Luo, and T. Newman. The CIECAM02 color appearance model. In *Color and Imaging Conference*, volume 2002, pages 23–27. Society for Imaging Science and Technology, 2002.

G. L. Murphy. 13 the contribution (and drawbacks) of models to the study of concepts. *Formal approaches in categorization*, page 299, 2011.

M. Nash and M. L. Donaldson. Word learning in children with vocabulary deficits. 2005.

J. O'keefe and L. Nadel. *The hippocampus as a cognitive map*. Oxford university press, 1978.

K. Patterson, P. J. Nestor, and T. T. Rogers. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature reviews neuroscience*, 8(12):976–987, 2007a.

K. Patterson, P. J. Nestor, and T. T. Rogers. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature reviews neuroscience*, 8(12):976–987, 2007b.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

M. V. Peelen and A. Caramazza. Conceptual object representations in human anterior temporal cortex. *Journal of Neuroscience*, 32(45):15728–15736, 2012.

G. Pengas, K. Patterson, R. J. Arnold, C. M. Bird, N. Burgess, and P. J. Nestor. Lost and found: bespoke memory testing for alzheimer's disease and semantic dementia. *Journal of Alzheimer's Disease*, 21(4):1347–1365, 2010.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

F. Pereira, S. Gershman, S. Ritter, and M. Botvinick. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive neuropsychology*, 33(3-4):175–190, 2016.

J. C. Peterson, D. Chen, and T. L. Griffiths. Parallelograms revisited: Exploring the limitations of vector space models for simple analogies. *Cognition*, 205:104440, 2020.

M. A. Pinheiro, J. Kybic, and P. Fua. Geometric graph matching using monte carlo tree search. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2171–2185, 2016.

G. Pobric, E. Jefferies, and M. A. L. Ralph. Category-specific versus category-general semantic impairment induced by transcranial magnetic stimulation. *Current biology*, 20(10):964–968, 2010.

S. F. Popham, A. G. Huth, N. Y. Bilenko, F. Deniz, J. S. Gao, A. O. Nunez-Elizalde, and J. L. Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, 24 (11):1628–1636, 2021.

M. C. Potter. Short-term conceptual memory for pictures. *Journal of experimental psychology: human learning and memory*, 2(5):509, 1976.

F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 2022.

W. V. O. Quine. *Word and object*. MIT press, 1960.

M. A. L. Ralph, E. Jefferies, K. Patterson, and T. T. Rogers. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1):42–55, 2017a.

M. A. L. Ralph, E. Jefferies, K. Patterson, and T. T. Rogers. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1):42–55, 2017b.

S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

E. Richardson and Y. Weiss. The surprising effectiveness of linear unsupervised image-to-image translation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7855–7861. IEEE, 2021.

L. E. Richland and N. Simms. Analogy, higher order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(2):177–192, 2015.

B. Riordan and M. N. Jones. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):303–345, 2011.

B. D. Roads and B. C. Love. Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, 2(1):76–82, 2020.

S. Robinson, J. Druks, J. Hodges, and P. Garrard. The treatment of object naming, definition, and object use in semantic dementia: The effectiveness of errorless learning. *Aphasiology*, 23(6):749–775, 2009.

T. T. Rogers, M. A. Lambon Ralph, P. Garrard, S. Bozeat, J. L. McClelland, J. R. Hodges, and K. Patterson. Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological review*, 111(1):205, 2004.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

B. C. Roy, M. C. Frank, P. DeCamp, M. Miller, and D. Roy. Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41):12663–12668, 2015.

D. K. Roy and A. P. Pentland. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146, 2002.

D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group. *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations*. MIT press, 1986.

Z. Sadeghi, J. L. McClelland, and P. Hoffman. You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, 76:52–61, 2015.

J. R. Saffran, R. N. Aslin, and E. L. Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996.

L. K. Samuelson, L. B. Smith, L. K. Perry, and J. P. Spencer. Grounding word learning in space. *PloS one*, 6(12):e28095, 2011.

M. P. Schadd, M. H. Winands, H. J. Van Den Herik, G. M.-B. Chaslot, and J. W. Uiterwijk. Single-player monte-carlo tree search. In *International Conference on Computers and Games*, pages 1–12. Springer, 2008.

T. Schnabel, I. Labutov, D. Mimno, and T. Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 298–307, 2015.

R. M. Schneider, D. Yurovsky, and M. Frank. Large-scale investigations of variability in children's first words. In *CogSci*, pages 2110–2115. Citeseer, 2015.

J. R. Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3 (3):417–424, 1980.

V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

R. Shao and D. Gentner. Perceptual alignment contributes to referential transparency in indirect learning. *Cognition*, 224:105061, 2022.

T. Sharot and C. R. Sunstein. How people decide what they want to know. *Nature Human Behaviour*, 4(1):14–19, 2020.

R. N. Shepard and S. Chipman. Second-order isomorphism of internal representations: Shapes of states. *Cognitive psychology*, 1(1):1–17, 1970.

R. Shetty, M. Rohrbach, L. Anne Hendricks, M. Fritz, and B. Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE international conference on computer vision*, pages 4135–4144, 2017.

L. A. Shneidman and S. Goldin-Meadow. Language input and acquisition in a mayan village: How important is directed speech? *Developmental science*, 15(5):659–673, 2012.

E. Shutova, D. Kiela, and J. Maillard. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 160–170, 2016.

G. A. Sigurdsson, J.-B. Alayrac, A. Nematzadeh, L. Smaira, M. Malinowski, J. Carreira, P. Blunsom, and A. Zisserman. Visual grounding in video for unsupervised word translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10850–10859, 2020.

C. Silberer and M. Lapata. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, 2012.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362 (6419):1140–1144, 2018.

J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, IEEE International Conference on*, volume 3, pages 1470–1470. IEEE Computer Society, 2003.

L. Smith and C. Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008.

L. B. Smith and L. K. Slone. A developmental approach to machine learning? *Frontiers in psychology*, page 2124, 2017.

J. S. Snowden and D. Neary. Relearning of verbal labels in semantic dementia. *Neuropsychologia*, 40(10):1715–1728, 2002.

R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26, 2013.

T. Stafford. How do we use computational models of cognitive processes? In *Connectionist models of neurocognition and emergent behavior: From theory to applications*, pages 326–342. World Scientific, 2012.

M. Stella, N. M. Beckage, and M. Brede. Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific reports*, 7(1):1–10, 2017.

M. Steyvers and J. B. Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78, 2005.

H. L. Storkel. Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of child language*, 36(2):291–321, 2009.

A. Suárez-González, C. G. Heredia, S. A. Savage, E. Gil-Néciga, N. García-Casares, E. Franco-Macías, M. L. Berthier, and D. Caine. Restoration of conceptual knowledge in a case of semantic dementia. *Neurocase*, 21(3): 309–321, 2015.

J. Sullivan, M. Mei, A. Perfors, E. Wojcik, and M. C. Frank. Saycam: A large, longitudinal audiovisual dataset recorded from the infants perspective. *Open mind*, 5:20–29, 2021.

C. S. Tamis-LeMonda, S. Custode, Y. Kuchirko, K. Escobar, and T. Lo. Routine language: Speech directed to infants during home activities. *Child development*, 90(6):2135–2152, 2019.

H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

S. Theves, G. Fernandez, and C. F. Doeller. The hippocampus encodes distances in multidimensional feature space. *Current Biology*, 29(7):1226–1231, 2019.

S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996.

V. Titouan, N. Courty, R. Tavenard, and R. Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pages 6275–6284. PMLR, 2019.

E. C. Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4): 189, 1948.

A. Tompary and S. L. Thompson-Schill. Semantic influences on episodic memory distortions. *Journal of Experimental Psychology: General*, 2021.

L. Unger, O. Savic, and V. M. Sloutsky. Statistical regularities shape semantic organization throughout development. *Cognition*, 198:104190, 2020a.

L. Unger, C. Vales, and A. V. Fisher. The role of co-occurrence statistics in developing semantic knowledge. *Cognitive Science*, 44(9):e12894, 2020b.

S. Van der Walt and N. Smith. A better default colormap for matplotlib [Conference session]. Python in Science (SciPy) Conference, Austin, TX, United States, 2015, July 6–12.

M. T. Van Kesteren, D. J. Ruiter, G. Fernández, and R. N. Henson. How schema and novelty augment memory formation. *Trends in neurosciences*, 35(4):211–219, 2012.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

P. Vetter, F. W. Smith, and L. Muckli. Decoding sound and imagery content in early visual cortex. *Current Biology*, 24(11):1256–1262, 2014.

P. Vetter, Ł. Bola, L. Reich, M. Bennett, L. Muckli, and A. Amedi. Decoding natural sounds in early visual cortex of congenitally blind individuals. *Current Biology*, 30(15):3039–3044, 2020.

G. Vigliocco, L. Meteyard, M. Andrews, and S. Kousta. Toward a theory of semantic representation. *Language and Cognition*, 1(2):219–247, 2009.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

M. Visser, K. Embleton, E. Jefferies, G. Parker, M. Lambon Ralph, et al. The anterior temporal lobes and semantic memory clarified: Novel evidence from distortion-corrected fmri. *Journal of Cognitive Neuroscience*, 6:1083–1094, 2010.

M. Visser, E. Jefferies, K. V. Embleton, and M. A. Lambon Ralph. Both the middle temporal gyrus and the ventral anterior temporal area are crucial for multimodal semantic processing: distortion-corrected fmri evidence for a double gradient of information convergence in the temporal lobes. *Journal of Cognitive Neuroscience*, 24(8):1766–1778, 2012.

R. Vogels and G. A. Orban. The effect of practice on the oblique effect in line orientation judgments. *Vision research*, 25(11):1679–1687, 1985.

Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

E. K. Warrington. The selective impairment of semantic memory. *The Quarterly journal of experimental psychology*, 27(4):635–657, 1975.

C. Wernicke. Wernickes works on aphasia: A sourcebook and review (pp. 91–145)[der aphasische symptomencomplex. eine psychologische studie auf anatomischer basis](gh eggert, trans.). *New York, NY: Mouton (Original work published 1874)*, 1977.

M. Won, J. Salamon, N. J. Bryan, G. J. Mysore, and X. Serra. Emotion embedding spaces for matching music to stories. *arXiv preprint arXiv:2111.13468*, 2021.

Y. Xian, B. Schiele, and Z. Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017.

C. Yu and L. B. Smith. Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, 18(5):414–420, 2007.

L. Zaadnoordijk, T. R. Besold, and R. Cusack. Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence*, 4(6):510–520, 2022.

M. Zhang, Y. Liu, H. Luan, and M. Sun. Earth movers distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, 2017.

Z.-H. Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.