A Conceptual Primitive Decomposition of the Sally-Anne Test

Jamie C. Macbeth[1], **Boming Zhang**[2], **Sharmin Badhan**[3]

[1]Department of Computer Science, Smith College, 10 Elm Street, Northampton, MA 01063 USA
[2]Manning College of Information and Computer Sciences, University of Massachusetts, Amherst, 140 Governors Drive, Amherst, Massachusetts 01003 USA
[3]Independent Researcher
jmacbeth@smith.edu, bomingzhang@umass.edu, sharminbadhan79@gmail.com

Abstract

Although large language models (LLMs) have been observed to perform at a human level in theory of mind tasks, deeper examinations and systematic testing of their performance in these domains is needed. Primitive decomposition representations show promise for building robotic systems with greater abilities for in-depth natural language understanding and generation. In this work, we explore representations of theory of mind which are combinations of conceptual primitives, focusing on simulations of a Sally-Anne false-belief test. We demonstrate how primitive decompositions into the conceptual building blocks of image schemas and Conceptual Dependency can represent the attribution of false beliefs to intelligent agents. The exploration has consequences for generating controlled and linguistically varied tests posed in natural language as challenge problems for large language models and for cognitive representations more broadly.

Introduction

Large language models (LLMs) have occasionally been observed to perform at or above humans level in certain kinds of theory of mind tasks. However, deeper examinations of LLM performance in highly controlled and varied environments (Mirzadeh et al. 2024) show the fragility of reported results and "highlight the importance of systematic testing to ensure a non-superficial comparison between human and artificial intelligences" (Strachan et al. 2024).

Primitive decomposition systems have, in the past, shown promise as representations for intelligent systems (Schank and Burstein 1982), and more recently they have been used to probe the understanding and reasoning capabilities of deep learning and large language models in the face of linguistic variation (Dai, Grandic, and Macbeth 2019). This work focuses on challenges of theory of mind representations in the image schemas and Minsky-Schank Conceptual Dependency Trans-frames tradition of nonverbal primitive decompositions. In this work, we simulate a Sally-Anne test to demonstrate how primitive decomposition systems, specifically decompositions into the conceptual building blocks of image schemas and Conceptual Dependency, can represent the attribution of false beliefs to intelligent agents. This work is part of a concerted effort to build a joint and unified image schemas Conceptual Dependency theory which is useful for linguistics and ontology theory and may also benefit the natural language understanding, story generation, inference, and reasoning capabilities of robotic applications (Lytinen 1992; Mandler and Cánovas 2014).

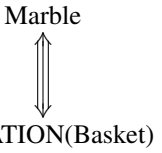

Figure 1: An example of a image schema Conceptual Dependency structure using the LOCATION primitive to indicate "The marble is in the basket." The triple arrow is used to indicate that a picture producer (in this case "Marble") can be described by a specified state (LOCATION) and a value for that state (Basket).

Primitive Decomposition

Image Schemas and Conceptual Dependency

A genre of work in cognitive artificial intelligence has recently juxtaposed and combined image schemas, preverbal conceptual building blocks developed by the cognitive linguistics and psychology communities (Mandler and Cánovas 2014), with the Conceptual Dependency Transframes system (CD), a traditional AI meaning representation evolved through AI systems for narrative understanding and generation (Macbeth, Gromann, and Hedblom 2017; Macbeth et al. 2023; Macbeth, Zhang, and Badhan 2025). We refer to this combined representation system as IS-CD.

The CD system (Schank 1975) is a theory for representing the meaning of natural language in a way that emphasizes the underlying conceptual levels rather than the surface syntax. At the conceptual level, CD represents meaning through conceptualizations, which are networks of interrelated concepts built from primitive categories such as PPs (*picture producers*), which typically denote entities or objects, PAs (*picture aiders*), which serve as modifiers that describe attributes of PPs, and ACTs (*actions*), which represent actions that can be performed or experienced. The relationships be-

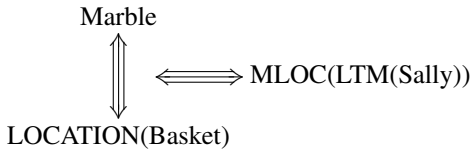

Figure 2: An example of a image schema Conceptual Dependency structure using the MLOC primitive. The horizontal triple arrow takes the "marble located in box" concept as an operand, and adds that that concept can be described by a specified state (MLOC) and a value (LTM(Sally), "Sally's long term memory"). A typical natural language gloss of this structure would be "Sally knows that the marble is in the basket".

tween these conceptual elements are expressed through dependency links that specify how parts of the concept depend on other parts for meaning.

CD represents meanings through structures that center on primitive acts. CD's six physical primitive ACTs—PTRANS, PROPEL, MOVE, INGEST, EXPEL, GRASP—have been mapped onto combinations of SOURCE_PATH_GOAL, SUPPORT, FORCE, CONTAINMENT, and PART-WHOLE image schemas (Macbeth, Gromann, and Hedblom 2017). CD structures with these primitives can have ACTOR, OBJECT, TO, and FROM conceptual *cases* which further specify the concept. Other result-focused (ATRANS), instrumental (SPEAK, ATTEND), and mental primitives (MTRANS, MBUILD) are used to represent acts of intelligence in the social world. By no means are these primitives a complete set for representing all possible meanings, and research into completing and improving the set of primitives is ongoing (Macbeth et al. 2023). But no well known image schemas are strong mappings for the mental act primitives of Conceptual Dependency which would be used to represent tests of theory of mind (Mandler and Cánovas 2014).

Mental ACT Primitives

The CD MTRANS primitive is typically defined as "the transfer of mental information from one individual to another," and it is used in decompositions of acts of speaking, reading, and other acts of language communication (Dyer 1982). When an MTRANS occurs between distinct parts of the same person's memory, it can represent notions of remembering, recalling, and forgetting. The CD MBUILD primitive represents "thought processes which create new conceptualizations from old ones" and acts described as "deciding," "realizing," "considering," and "imagining" (Dyer 1982).

While CD conceptualizations with physical primitives such as PTRANS/SOURCE_PATH_GOAL take physical objects as their OBJECT cases, MBUILD and MTRANS take other CD conceptualizations as their MOBJECTs, or "mental objects". Additionally, conceptualizations with physical primitives such as PTRANS/SOURCE_PATH_GOAL can take physical objects in their TO or FROM cases to indicate the origin and destination of an object's movement, while MBUILD and MTRANS take certain classes of locations where an MOBJECT can go or be created. Conceptual Dependency specifies that intelligent beings have *long term memory* (LTM), *intermediate memory* (IM), and a *conceptual processor* (CP) as locations for MOBJECTs, with the CP being the place for conscious thought (Schank 1975).

False Beliefs in IS-CD

Static State Representations

The identification of MBUILD and MTRANS ACT primitives reflects the way that Conceptual Dependency was initially used to represent acts and events in stories, where the focus in story understanding systems is on tracking changes in the state of the world, rather than "static" situations or states of being. Conceptual Dependency additionally has CONTAIN, PART, and LOC/LOCATION primitive elements to represent static relationships and dependencies between physical objects. These elements map onto CONTAINMENT, PART-WHOLE, and LOCATION image schemas (Mandler and Cánovas 2014). Figure 1 shows an example CD structure using LOC/LOCATION to form a conceptualization about a static state of affairs, that the marble is in the basket. The CD structure appears as a diagram with a triple arrow between the picture producer "marble" and the picture aider LOCATION(Basket), which indicates the relationship that LOCATION(Basket) describes is an attribute of "marble" or, equivalently, that "basket" and "marble" have a LOCATION relationship.

The MLOC "Mental Location" Primitive

Conceptual Dependency additionally has a primitive element MLOC, a conceptual attribute for mental location, which has no counterpart in image schemas. Figure 2 shows an example of an image schema Conceptual Dependency structure which represents a static situation of belief using MLOC. The horizontal triple arrow takes the "marble located in basket" concept as an operand, and represents that that concept can be described by a specified state, MLOC, and a value, LTM(Sally), "Sally's long term memory". This indicates that the "marble located in the basket" concept is located in Sally's LTM. A typical natural language gloss of this structure would be "Sally knows that the marble is in the basket".

False Belief Structures in IS-CD

The Sally-Anne test is a psychological experiment used to assess theory of mind (Baron-Cohen, Leslie, and Frith 1985). The person taking the test watches two characters, Sally and Anne, interact with an object, typically a marble. Sally places the marble in a basket and leaves the scene; Anne then moves the marble to a box. The person is asked where Sally will look for the marble upon returning, revealing their understanding that Sally holds a false belief.

In the CD formalism (Schank 1975) all conceptualizations in an intelligent being's LTM are believed or held to be true by that being. To represent that an intelligent being "knows

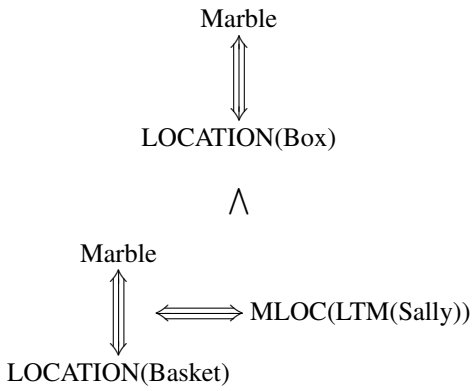

Figure 3: The false belief structure as a image schema Conceptual Dependency structure for Sally in the classic Sally-Anne test, moments after Anne has moved the marble from the basket to the box without Sally observing the move. A structure representing a false belief is a concept in an intelligent being's LTM that turns out to be false in the real world. This IS-CD structure uses CD's logical AND operator to combine the structures from Figures 1 and 2.

something" is done by having the conceptualization with the MLOC attribute. A concept in an intelligent being's LTM that turns out to be false in the real world is a false belief. Therefore we are able to construct a false belief structure for Sally in the classic Sally-Anne test, moments after Anne has moved the marble from the basket to the box without Sally observing the move. Figure 3 shows this IS-CD structure, which uses CD's logical AND operator to combine the structures from Figures 1 and 2.

One important aspect of this structure is that it is focused on the point of view of an understander system which stands apart from Sally and Anne and has observed or has processed the Sally-Anne test "story". The structure contains the understander's conceptualization of the truth the way it has understood it, that the marble is in the box. It additionally contains the MLOC structure representing that the conceptualization that the marble is in the basket is in Sally's long-term memory, and thus she falsely believes it. When an understanding system performs a surface realization of the full conceptualization, it will need to take into account the conflict between the understood fact and the false belief to produce "Sally thinks that the marble is in the basket" instead of "Sally knows that the marble is in the basket". Typically the question posed to humans in the Sally-Anne test is about where Sally will look for the marble. Determining ways to representing complex intelligent behaviors such as "looking" or "searching" in IS-CD is an open area of research.

Conclusion

This brief paper presents an attempt to show how a false-belief can be represented within the IS-CD framework. What we find most important in this exploration was that the primitive decomposition system has to be able to separate the real world from the world of thoughts and imagery. Future work may use natural language understanding and paraphrasing systems that have been established for CD to demonstrate understanding of the test, to integrate these understanding systems into robotics, and to generate a controlled variety of false-belief tests posed in natural language as challenge problems for large language models.

References

Baron-Cohen, S.; Leslie, A. M.; and Frith, U. 1985. Does the Autistic Child Have a "Theory of Mind"? *Cognition*, 21(1): 37–46.

Dai, M.; Grandic, S.; and Macbeth, J. C. 2019. Linguistic Variation and Anomalies in Comparisons of Human and Machine-Generated Image Captions. *Advances in Cognitive Systems*, 8: 33–51.

Dyer, M. G. 1982. *In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension*. Cambridge, MA: MIT Press.

Lytinen, S. L. 1992. Conceptual dependency and its descendants. *Computers & Mathematics with Applications*, 23(2): 51–73.

Macbeth, J. C.; Gromann, D.; and Hedblom, M. M. 2017. Image Schemas and Conceptual Dependency Primitives: A Comparison. In *Proceedings of The Joint Ontology Workshops, Episode 3: The Tyrolean Autumn of Ontology*. Bolzano-Bozen, Italy: The International Association for Ontology and its Applications.

Macbeth, J. C.; Kilayko, A.; Zhao, Z.; Song, S.; and Zheng, W. X. 2023. Image Schema Decompositions of the Conceptual Dependency INGEST Primitive: A Study of Paraphrases. In *Proceedings of The Seventh Image Schema Day (ISD7)*. Rhodes, Greece: The International Association for Ontology and its Applications.

Macbeth, J. C.; Zhang, B.; and Badhan, S. 2025. Script-based Inferences in an Image Schema Story Understander. In *Proceedings of The Ninth Image Schema Day (ISD9)*. Cantania, Italy: The International Association for Ontology and its Applications.

Mandler, J. M.; and Cánovas, C. P. 2014. On Defining Image Schemas. *Language and Cognition*, 6(4): 510–532.

Mirzadeh, I.; Alizadeh, K.; Shahrokhi, H.; Tuzel, O.; Bengio, S.; and Farajtabar, M. 2024. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. arXiv:2410.05229.

Schank, R. C. 1975. *Conceptual Information Processing*. New York, NY: Elsevier.

Schank, R. C.; and Burstein, M. 1982. Modeling Memory for Language Understanding. Research Report #220, Yale University, Department of Computer Science, New Haven, CT.

Strachan, J. W.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7): 1285–1295.