**Cognitive Processing manuscript No.**
(will be inserted by the editor)

# Human inference beyond syllogisms: an approach using external graphical representations

**Yuri Sato · Gem Stapleton · Mateja Jamnik · Zohreh Shams**

**Abstract** Research in psychology about reasoning has often been restricted to relatively inexpressive statements involving quantifiers (e.g., syllogisms). This is limited to situations that typically do not arise in practical settings, like ontology engineering. In order to provide an analysis of *inference*, we focus on reasoning tasks presented in external graphic representations where statements correspond to those involving multiple quantifiers and unary and binary relations. Our experiment measured participants' performance when reasoning with two notations. The first notation used topological constraints to convey information via node-link diagrams (i.e., graphs). The second used topological *and* spatial constraints to convey information (Euler diagrams with additional graph-like syntax). We found that topological-spatial representations were more effective for inferences than topological representations alone. Reasoning with statements involving multiple quantifiers was harder than reasoning with single quantifiers in topological representations, but not in topological-spatial representations. These findings are compared to those in sentential reasoning tasks.

Y. Sato (corresponding author)
Centre for Secure, Intelligent and Usable Systems, University of Brighton, UK
E-mail: y.sato@brighton.ac.uk

G. Stapleton
Centre for Secure, Intelligent and Usable Systems, University of Brighton, UK
E-mail: g.e.stapleton@brighton.ac.uk

M. Jamnik
Department of Computer Science and Technology, University of Cambridge, UK
E-mail: mateja.jamnik@cl.cam.ac.uk

Z. Shams
Department of Computer Science and Technology, University of Cambridge, UK
E-mail: zohreh.shams@cl.cam.ac.uk

**Keywords** inference · diagrammatic reasoning · external representation · quantifiers · binary predicates

## 1 Introduction

In the literature on psychology of reasoning, categorical syllogisms, for example, *All B are A; some B are C*; therefore *some C are A*, have been intensively studied (for a survey, see Khemlani & Johnson-Laird 2012). However, it may not be true that categorical syllogisms are frequently used in our daily life. Non-syllogistic forms of reasoning have attracted particular attention in the study presented in this paper, which encompasses the expressive case of binary verbs (requiring two terms) and multiple quantifiers, such as *All koalas eat only eucalyptuses*. This is achieved through external graphical representations where quantification is implicit in the formed statements; examples will be given later.

Some cognitive studies have already explored frameworks that are beyond the traditional ones of categorical syllogisms. Johnson-Laird, Byrne, and Tabossi (1989) dealt with syllogisms involving verbs, for example, from the premises *All A are in the same place as some B; All B are in the same place as all C* to the conclusion *All A are in the same place as all C*. However, the verbs are restricted to spatial binary ones with transitivity or symmetry. A similar restriction was applied to the recent study of Ragni and Sonntag (2012). These studies adopt an approach from the viewpoint of mental model theory, and propose that mental representations for multiply quantified sentences consist simply of alphabets or dots with different shapes representing individuals. However, such simple representations, by way of only using spatial verbs, do not represent all cases that use binary verbs.

Geurts and van der Slik (2005) used mixed forms of syllogisms with single and double quantifiers; for example, *Most A played against more than two B* and *All B were C* implies *Most A played against more than two C*. They demonstrated the effect of monotonicity profiles of quantifiers, rather than specific mental representations and processes. In addition, non-standard quantifiers, for example, cardinal (numerical) quantifiers such as *more than three* (Kroger et al., 2008), proper nouns such as *a is an A* ("a" is an individual constant) (Politzer & Mercier, 2008; Khemlani, Lotstein & Johnson-Laird, 2014), and proportional quantifiers such as *most* and *few* (Ragni, Singmann, & Steinlein, 2014; Sato & Mineshima, 2016) have been also explored. However, the scope of the studies was still restricted to the syllogistic form consisting of minor, major, and middle terms.[1] Furthermore, each extended case was explored separately and there have been few comparisons between single quantifiers and multiple quantifiers.

By contrast, recent developments in ontology engineering shed new light on natural language inferences, contributing to the expanding coverage of psychological reasoning tasks. Nguyen et al. (2012) collected deduction patterns (inference tasks), and demonstrated their prevalence of a wide variety of forms of reasoning with quantifiers and unary and binary relations. Using novice participants, performance using the deduction patterns as inference

[1] Middle terms appear in both major premises (containing major terms) and minor premises (containing minor terms), and minor and major terms compose conclusions.

tasks was evaluated. It was found that there was a difficulty gap between reasoning with single quantifiers and reasoning with multiple quantifiers, with 76.4% accuracy for (statements involving) single quantifiers and 59.0% for multiple quantifiers. This leads to two important questions. What makes reasoning with multiple quantifiers hard? What kind of cognitive processes of reasoning do people employ in the presence of quantifiers?

A main concern of these questions is *inference tasks*, which are distinguished from interpretations, but inference necessarily follows the process of interpreting premises. So it is important for the interpretation to be fixed in some way when exploring the nature of human inferences. Indeed, Stenning and van Lambalgen (2001) emphasised the distinction of two kinds of reasoning: reasoning *toward* an interpretation of premises and reasoning *from* a fixed interpretation of premises.[2] Based on the approach of using easy-to-understand representations (Sato & Mineshima, 2015; Sato, Sugimoto, & Ueda, 2018), the current study adapts external graphical representations, instead of ordinary representations of natural language sentences, to fix the interpretation of premises and provide a fine-grained analysis of inference. In particular, we focus on two distinct graphical representations. The first exploits topology to convey information via node-link diagrams (i.e., graphs), hereinafter called a *topological representation*. The second exploits topological *and* spatial constraints to convey information (Euler diagrams with additional graph-like syntax), hereinafter called a *topo-spatial representation*. Importantly, both types of representation convey information that is semantically rich. They give rise to statements involving either single quantifiers (and unary relations) or involving multiple quantifiers along with binary relations.

In this paper, we explore the cognitive processes involved in graphical inferences about semantically rich statements containing multiply quantified information. Of course, the findings apply to the cases covered by the particular representations used in the study. However, they lead us to a more general point. Even when people interpret statements of sentences in usual sentential reasoning tasks, choices must be made about their representations. Thus, the findings in our experiment, in which particular representations are externally given, can contribute to our understanding of human inferences in sentential format. This will be explored later in the paper. The structure of this paper is as follows. In Section 2, we present a task analysis of deductive inference by graphical representations. In Section 3, we report on the experiment which measured participants' performance of the tasks. In Section 4, we discuss the findings and implications to cognitive science of reasoning.

## 2 Task analysis

### 2.1 Topological and topo-spatial representations

As mentioned in the Introduction, human reasoning with quantifiers is necessary in a wide variety of single and multiple quantified forms, in line with the recent development in ontology

[2] In addition, Stenning and van Lambalgen clearly stated that "If the interpretation is not fixed, what is one actually testing? (p.291)".

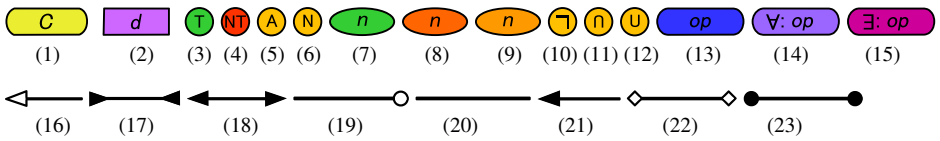


**Fig. 1** Syntactic units constituting SOVA graphs: (1) classes, (2) individuals, (3) all things, (4) no things, (5) anonymous classes, (6) natural numbers, (7–9) minimum/maximum/exact cardinalities, (10) negation, (11) intersection, (12) union, (13) binary-verbs, (14) binary-verbs with universal restriction, (15) binary-verbs with existential restriction, (16) isSubclassOf, (17) DisjointClasses, (18) EquivalentClasses, (19) IntersectionOf in linking ∩, UnionOf in linking ∪, ComplementOf in linking ¬, (20) InstanceOf in linking between class and individual, (21) binary-verb relation, (22) binary-verb relation with universal restriction and "everything" in subject term, (23) binary-verb relation with universal restriction and "everything" in object term (concrete words are inserted in nodes with italic font labels).

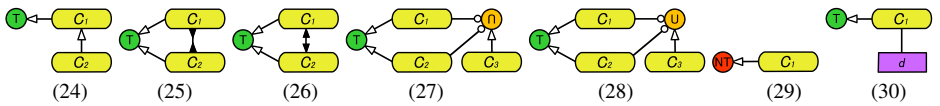


**Fig. 2** Basic configurations of SOVA graphs: (24) $C_1$ is a subset of $C_2$. (25) $C_1$ is disjoint from $C_2$. (26) $C_1$ is equivalent to $C_2$. (27) $C_3$ is a subset of the intersection of $C_1$ and $C_2$. (28) $C_3$ is a subset of the union of $C_1$ and $C_2$. (29) $C_1$ is a subset of nothing; i.e., $C_1$ is empty. (30) $d$ is an instance of $C_1$.

engineering. Historically, ontologies were defined using some kind of graphs (e.g., Brachman & Schmolze, 1985), before the formulations of description logics (Schmidt-Schauß & Smolka, 1991) and web ontology languages (W3C OWL Working Group, 2012).[3] As a graph representation for OWL constructs, this study focuses on the Simple Ontology Visualization API: SOVA[4] (Itzik & Reinhartz-Berger, 2014). Note that SOVA graphs are fundamentally composed of nodes and links but, as full ontology representations, they need a wider variety of nodes and links.

The basic syntactic units of SOVA graphs are given in Fig.1. Nodes with oblique font labels are used for named classes (i.e., sets), individuals, cardinalities, binary-verbs, etc. In nodes for binary-verbs, there are two kinds of restrictions: (1) in universal restriction (∀:) – subject class *op only* object class, where *op* is a binary verb; (2) in existential restriction (∃:) – subject class *op at least* one object class. Nodes with normal font labels are used for all things (T), no things (NT), anonymous classes (A), negation (¬), intersection (∩), etc. In addition, various forms of links are used for two types of relations: set-theoretical ones (e.g., SubclassOf, DisjointClasses, IntersectionOf) and binary-verb. Fig.2 illustrates examples of basic configurations of SOVA graphs. Examples with binary verbs of SOVA graphs will be given in Section 2.2. SOVA graphs, whilst more complex than those in Hartley and Barnden (1997, p.170), are still composed of topological relationships in nodes and links, giving rise to semantic meanings of expressive statements involving quantifiers.

[3] As discussed in Stenning (2002, Chap.2), graph (node-link) representations can essentially be the same as sentential representations with respect to expressivity.

[4] `https://protegewiki.stanford.edu/wiki/SOVA` (Accessed Dec. 2017)

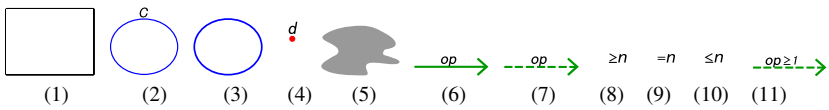


**Fig. 3** Syntactic elements constituting concept diagrams: (1) all things, (2) named classes, (3) anonymous classes, (4) named individuals, (5) the absence of things, (6) binary-verbs with universal restriction, (7) binary-verbs with "at least" constraints, (8–10) cardinal restrictions which are written on arrows, (11) example of a cardinal restriction: binary-verbs with existential restriction.

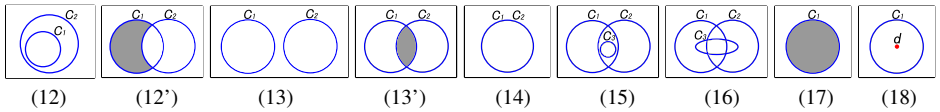


**Fig. 4** Basic configurations of concept diagrams: (12–12') $C_1$ is a subset of $C_2$. (13–13') $C_1$ is disjoint from $C_2$. (14) $C_1$ is equivalent to $C_2$. (15) $C_3$ is a subset of the intersection of $C_1$ and $C_2$. (16) $C_3$ is a subset of the union of $C_1$ and $C_2$. (17) $C_1$ is empty. (18) $d$ is an instance of $C_1$.

As well as topological relations, spatial relations – such as inclusion and exclusion – are available. We focus on concept diagrams (Stapleton, Compton, & Howse, 2017) as a topo-spatial representation of ontologies, as shown in Fig.3. The basic idea of concept diagrams is that Euler diagrams and graphs (nodes and links) are merged; for a similar approach, see Harel (1988) and Sugiyama and Misue (1991), although those systems are not expressive enough to fully represent ontologies. Syntactic units constituting concept diagrams are given in Fig.3. Rectangles are used for all things, named (labelled) curves are used for classes, unnamed curves for anonymous classes, and shading illustrates the absence of things. Solid arrows with labels are used for binary-verbs with universal (only) restriction, and dashed arrows with labels are used for binary-verbs with existential (at least one) restriction. Furthermore, relations are divided into two types: set-theoretical and binary-verb. Set-theoretical relations (e.g., SubclassOf, DisjointClasses, IntersectionOf, etc.) are expressed by spatial (inclusion and exclusion) relations of the corresponding syntactic objects; see Fig.4. Binary-verb relations are expressed by arrows from source objects to target objects. Examples with binary verbs of concept diagrams will be given in Section 2.2.

## 2.2 Reasoning with multiple quantifiers in topological and topo-spatial representations

We can make inferences using topological representations of SOVA graphs and using topo-spatial representations of concept diagrams. In such diagrammatic reasoning tasks, one is asked to judge whether the diagram transformations from premises to a conclusion are valid. Here, the premise diagrams are true and they are transformed into the conclusion diagram. If the conclusion diagram is true, given the information in the premise diagrams, the transformation is valid. Otherwise the transformation is not valid.

Fig.5 shows examples of tasks in graphical reasoning, which consist of two premises and one conclusion, and they are divided by a line. In each figure, (a) and (b) are SOVA cases (c) and (d) are concept diagram cases. Cases (a) and (c) give rise to statements about sets and unary relations involving single quantifiers, hereinafter called the 'single quantifier' case, and are translated as *Everything is a darfellan & No grippli is a darfellan; therefore, nothing is a grippli*. Cases (b) and (d) give rise to the statements about sets and binary relations involving multiple (double) quantifiers, hereinafter called the 'multiple quantifier' case, and are translated as *Every daemonfey is related to at least one thing in both axani and phoera under 'isGuidedBy' & No axani is a phoera; therefore, nothing is a daemonfey*.

In graphical or diagrammatic reasoning, it is generally assumed that reasoners merge premise representations into one, and then judge if a given conclusion representation can be obtained from the merged representation (cf. Stenning & Oberlander, 1995). Fig.6 illustrates possible intermediate representations generated by merging premise diagrams in Fig.5. In the SOVA case of Fig.5(a), one identifies the nodes T and darfellan in the two premise graphs, and overlays the two linking (double headed and white headed arrows) between these two nodes, while preserving the other relations. Consequently, one can produce the merged representation of Fig.6(a). Here T, referring to all-things, is linked with an equivalent node to darfellan, while T is also linked by a subclass arrow from grippli (which is disjoint from darfellan). In order for all the premises and conclusion to be true, it is needed as a conclusion statement that there is no individual who is a grippli; that is, the set grippli is empty. How do we deduce this from the SOVA graph in Fig.6(a)? Well, Fig.6(a) explicitly shows that darfellan is equivalent to Thing and is disjoint from grippli. Therefore, grippli is disjoint from Thing and, so, there can be no grippli individuals. Thus, one can judge that the conclusion statement corresponding to *Nothing is a grippli*, which is expressed in the graph by linking, using a subclass arrow from grippli to NT (which represents the empty set), is valid. In the concept diagram case of Fig.5(c), Fig.6(c) is created by identifying the curves of darfellan in the two premise diagrams and merging the diagrams. Here the information that the region other than darfellan is shaded is added into the diagram in which the intersection of grippli and darfellan is shaded. As a result, the region inside of darfellan, but outside of grippli is unshaded. From this merged diagram, we can extract the conclusion diagram in which the region inside grippli is entirely shaded. Again, we deduce there can be no grippli individuals. Here we can observe that, in reasoning with single quantifiers, SOVA graph and concept diagrams are similar in that unary relations are expressed by one basic configuration: SOVA links two nodes and concept diagrams use spatial relations.

Fig.6(b) is created by identifying the nodes T, axani, and phoera in the two premise diagrams of Fig.5(b) and merging the diagrams. While daemonfey is linked under a verb to at least one thing in the intersection between axani and phoera, axani is linked using a disjoint edge to phoera. In order for all the premises and conclusion to be true, the conclusion statement in which daemonfey is empty is needed. Thus, one can judge that the conclusion statement corresponding to *Nothing is a daemonfey* is valid. Here, understanding the relations between daemonfey and axani/phoera through a binary verb requires processing *multiple* arrows among nodes. Thus, reasoning with multiple quantifiers is expected to require much more cognitive effort than reasoning with single quantifiers. On the other hand, in concept

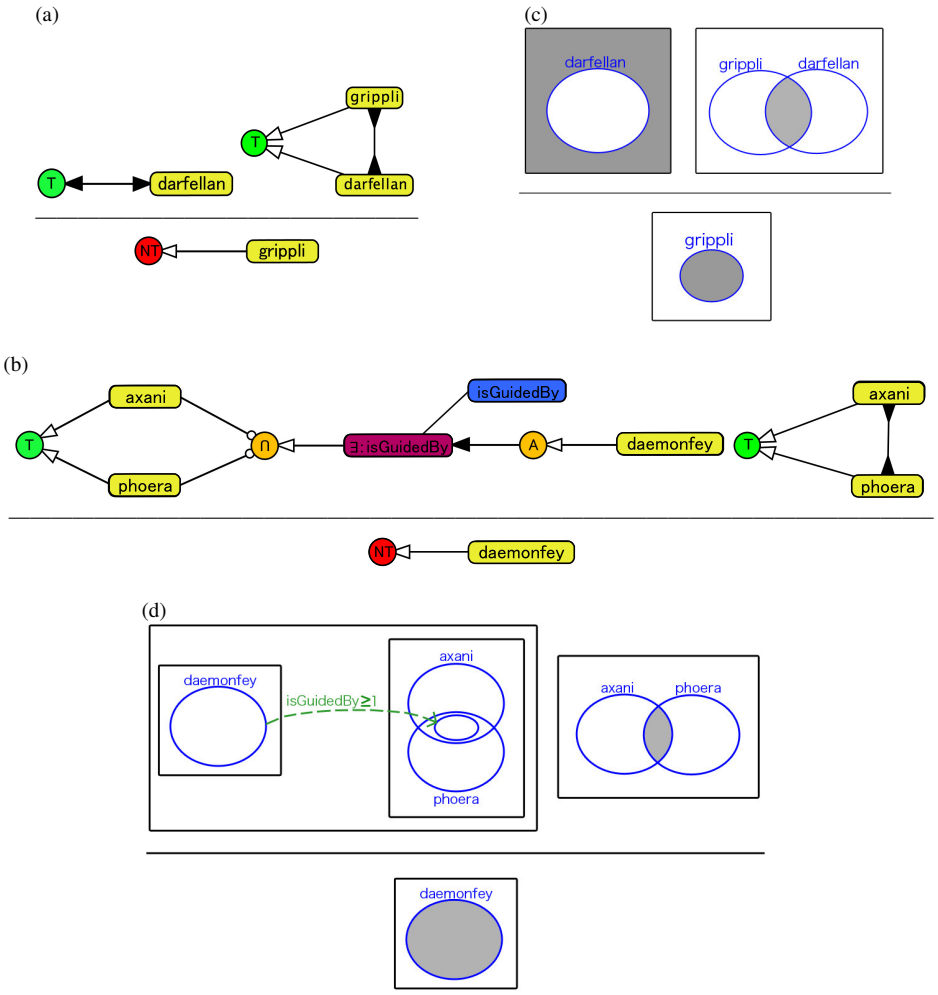


**Fig. 5** Task examples. Two premises and one conclusion are divided by a line: (a) topological representations in single quantifier case; (b) topological representations in multiple quantifier case; (c) topo-spatial representations in single quantifier case; (d) topo-spatial representations in multiple quantifier case.

diagrams, as shown in Fig.6(d), binary verbs are expressed by one arrow in which the source and target are directly specified (since unary relations are expressed by spatial constraints instead of arrows). Thus, reasoning with multiple quantifiers may not require much more effort than reasoning with single quantifiers.

In some cases of reasoning with concept diagrams, there might be certain difficulties in merging premise diagrams. As an example, consider the task shown in Fig.7. Judging

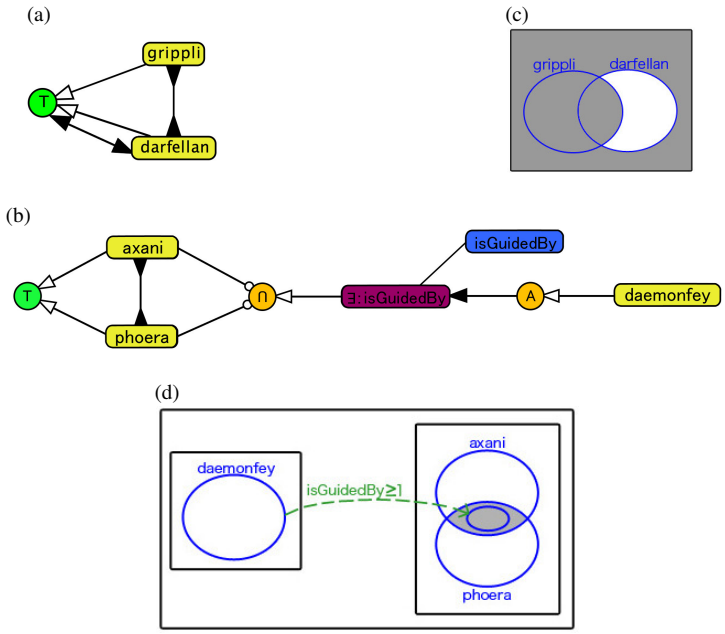


**Fig. 6** Possible intermediate representations by merging premise diagrams, corresponding to the cases (a)–(d) in Fig.5.

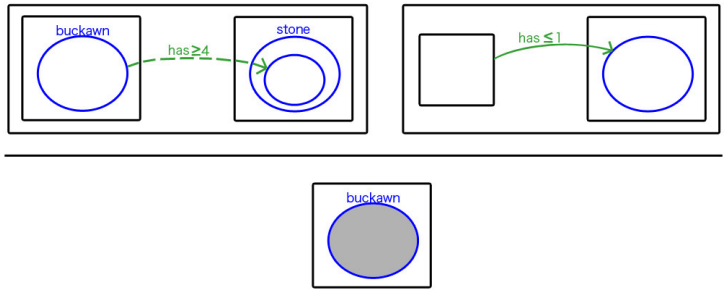


**Fig. 7** A task example of 'non-matching case' in topo-spatial representations: premise diagrams do not match, but can still be merged.

from curve configurations, neither inclusion nor exclusion relations necessarily hold between the (stone and unlabelled) curves in the first premise and an unlabelled curve in the second premise.[5] Judging from binary verb relations, however, since all things (rectangle) are related

[5] It should be noted that throughout the paper, each class (concept) may denote an empty or non-empty set. Consequently, each minimal region (i.e., a region having no other region contained within it) in a concept diagram also may denote an empty or a non-empty set. Thus, when the relation between two curves is unknown, they

to only a class (curve) under *has* (in the second premise), one can find that a subclass of things (curve labelled buckawn) cannot be related to something other than the above class under *has*. Thus, in either possible configuration between the stone curve in the first premise and the unlabelled curve in the second premise, the target region of the arrow 'has at least four' is inside the target region of the arrow 'has at most one'. This expresses that *each buckawn has at least four stones* and *each buckawn has at most one stone*. Accordingly, individuals in such a subject class (buckawn) cannot exist. In order for all the premises and conclusion to be true, the conclusion statement in which the subject class buckawn is empty is needed. Thus, one can judge that the conclusion statement corresponding to *Nothing is a buckawn* is valid. Cases in which premise diagrams do not uniquely match when merging them are relatively complex in that the reasoner has to consider more than one possibility; we call them 'non-matching' cases.[6] Solving non-matching cases like Fig.7 could involve processes requiring more cognitive effort than 'matching' cases like Fig.5(c-d). In reasoning with topological representations of SOVA, there is no such classification in merging premise representations.

Based on our analysis, we have the following five hypotheses. (1) In reasoning with single quantifiers, there is no significant difference between the two representations. (2) Topo-spatial representations are more effective than topological representations in reasoning with multiple quantifiers. (3) In topological representations, reasoning with multiple quantifiers is harder than reasoning with single quantifiers. (4) In topo-spatial representations, there is no significant difference between single and multiple quantifiers. (5) In topo-spatial representations, reasoning in the non-matching cases is harder than reasoning in the matching cases.[7]

## 3 Experiment

In order to test the above five hypotheses, we conducted an experiment measuring participants' performances when making inferences with topological representations and topo-spatial representations. Before the inference tasks, the participants were given instructions on the (informal) semantics of representations and then their comprehensions were checked by

are represented as two partially overlapping curves. Note that full understanding of diagrammatic semantics with respect to partially overlapping curves is not required for understanding the cognitive strategy described here. If people find that an inclusion or exclusion relationship is unspecified to be merged, they can judge the validities of arguments. Indeed, in Sato, Wajima, and Ueda (2018) who recorded the coordinate values of diagrams as they were moved by reasoners, it was reported that some reasoners really enumerated multiple possible configurations of inclusion, exclusion, and overlapping relationships, instead of placing curves as partially overlapping.

[6] This kind of difficulty in premise integration can also arise in reasoning with Euler diagrams in cases of syllogisms having an existential premise (e.g., *All B are A; some C are B. Therefore, some C are A*); see Sato and Mineshima (2015).

[7] In statistical hypothesis testing, the null hypotheses corresponding to the above predictions are that there is no significant difference between the two conditions. While the predictions of (2), (3), and (5) indicate that the null hypotheses are expected to be rejected, the predictions of (1) and (4) indicate that the null hypotheses are not expected to be rejected. Of course, there is no rigorous method of validating that no statistically significant difference (effect) exists. However, should we find no significant difference then this, at the very least, supports our predictions (1) and (4) and, thus, allows us to discuss the nature of human inference in question here.

a pretest (for details, see below). Given the experimental settings, it is assumed in our framework that the tasks for participants involve not just syntactic transformations of the pieces of some plane figure puzzle, but a kind of *diagrammatic inference*, in which the entailment from premise to a conclusion is judged using both syntactic and semantic information of diagrammatic or graphical representations.[8]

## 3.1 Method

### *3.1.1 Participants*

Forty-five undergraduate students from classes on elementary computer science at the University of Brighton were recruited. The mean age was 22.53 ($SD = 5.92$) with a range of 18–48 years. All participants gave informed consent and were paid for their participation (by way of refreshment voucher). The experiment method was approved by the CEM School Research Ethics Panel of the University of Brighton. None had any prior training of ontology engineering or syllogistic logic. One participant gave up on the way, and their data was excluded. Participants were divided into two groups: the topological group ($N = 19$) and the topo-spatial group ($N = 25$).

### *3.1.2 Materials*

The participants in each group were presented with premise graphs/diagrams and a conclusion graph/diagram (such as Fig.5). Participants were asked to answer the question of whether the graph/diagram transformations from premises to conclusion were valid. As shown in Table 1, we presented 20 items in total, out of which 10 items consisted only of validly transformed diagrams (#01–10) and 10 items included invalidly transformed diagrams (#11–20). The 10 valid items were selected from the medium difficulty patterns in Nguyen et al. (2012). The invalid items were created from the valid ones with minimal changes. Furthermore, the tasks were divided into singly quantified cases (#01–05/#11–15) and multiply quantified cases (#06–10/#16–20), and divided into matching cases (#1, 2, 4, 5, 7, 9, 11, 14, 15, 17, 19) and in non-matching cases (#3, 6, 8, 10, 12, 13, 16, 18, 20). The tasks were presented in one of three random orders and as a paper-and-pencil test. There was no time limit for completing the tasks, although the approximate time (30 minutes) for taking the experiment was suggested to the participants.

[8] This is consistent with the view that diagrammatic reasoning is a kind of *surrogated* reasoning (Barwise & Shimojima, 1995), in which reasoning about statements having semantic values is partially taken over by operations on external aids such as diagrams (for various examples of diagrammatic reasoning, see Glasgow, Narayanan & Chandrasekaran, 1995). Notions related to surrogation have also been discussed as analogical mapping in Gentner, Holyoak and Kokinov (2001), Funt (1980), Steels (1990), and Chandrasekaran (2011).

**Table 1** Lists of experimental tasks and results of accuracies. Upper rows are described in a natural language and lower rows are described in description logic. #01–10 are valid items; #11–20 are invalid items. #01–05/#11–15 are relevant to single quantifiers only; #06–10/#16–20 are relevant to multiple quantifiers.

| No | Premises ⇒ Conclusion | Topo.% | TopoSpa.% |
|---|---|---|---|
| 01. | (Every $A$ is a $B$) & (no $A$ is a $B$) ⇒ (nothing is an $A$) | | |
| | $(A \sqsubseteq B) \wedge (Dis(A,B)) \Rightarrow (A \sqsubseteq \bot)$ | 62.5 | 80.1 |
| 02. | (No $A$ is a $B$) & (every $C$ is an $A$) & (every $D$ is a $B$) ⇒ (no $C$ is a $D$) | | |
| | $(Dis(A,B)) \wedge (C \sqsubseteq A) \wedge (D \sqsubseteq B) \Rightarrow (Dis(C,D))$ | 75.0 | 85.7 |
| 03. | (Every $A$ is a (every $B$ is a $C$)) & (every $B$ is a $C$) ⇒ (every $A$ is a $C$) | | |
| | $(A \sqsubseteq (B \sqsubseteq C)) \wedge (B \sqsubseteq C) \Rightarrow (A \sqsubseteq C)$ | 87.5 | 85.7 |
| 04. | (Everything is a $B$) & (no $A$ is a $B$) ⇒ (nothing is an $A$) | | |
| | $(\top \sqsubseteq B) \wedge (Dis(A,B)) \Rightarrow (A \sqsubseteq \bot)$ | 56.3 | 66.7 |
| 05. | (Every $A$ is a $B$) & (every $A$ is non-$B$) ⇒ (nothing is an $A$) | | |
| | $(A \sqsubseteq B) \wedge (A \sqsubseteq \neg B) \Rightarrow (A \sqsubseteq \bot)$ | 50.0 | 76.2 |
| 06. | (Every $A$ is related to at least one $B$ under $R$) & (Everything that something is related to under $R$ is a $C$) ⇒ (every $C$ is related to at least one thing in both $B$ and $C$) | | |
| | $(A \sqsubseteq \exists R.B) \wedge (Rang(R.,C)) \Rightarrow (C \sqsubseteq \exists R.(B \sqcap C))$ | 87.5 | 80.1 |
| 07. | (Every $A$ is related to at least one thing in both $B$ and $C$ under $R$) & (no $B$ is a $C$) ⇒ (nothing is an $A$) | | |
| | $(A \sqsubseteq \exists R.(B \sqcap C)) \wedge (Dis(B,C)) \Rightarrow (A \sqsubseteq \bot)$ | 25.0 | 85.7 |
| 08. | (Every $A$ is related to at least three things in $B$ under $R$) & (every $A$ is related to at most one $B$ under $R$) ⇒ (nothing is an $A$) | | |
| | $(A \sqsubseteq \geq 3R.B) \wedge (A \sqsubseteq \leq 1R.B) \Rightarrow (A \sqsubseteq \bot)$ | 43.8 | 52.4 |
| 09. | (Every $A$ is related to at least one $B$ under $R$) & (every $B$ is nothing) ⇒ (nothing is an $A$) | | |
| | $(A \sqsubseteq \exists R.B) \wedge (B \sqsubseteq \bot) \Rightarrow (A \sqsubseteq \bot)$ | 62.5 | 57.1 |
| 10. | (Every $A$ is related to at least four things in $B$ under $R$) & (each thing is related to at most one thing under $R$) ⇒ (nothing is an $A$) | | |
| | $(A \sqsubseteq \geq 4R.B) \wedge (Fun(R.)) \Rightarrow (A \sqsubseteq \bot)$ | 31.3 | 47.6 |
| 11. | (Every $B$ is a $A$) & (no $A$ is a $B$) ⇒ (nothing is an $A$) | | |
| | $(B \sqsubseteq A) \wedge (Dis(A,B)) \Rightarrow (A \sqsubseteq \bot)$ | 75.0 | 71.4 |
| 12. | (No $A$ is a $B$) & (every $C$ is an $A$) & (every $B$ is a $D$) ⇒ (no $C$ is a $D$) | | |
| | $(Dis(A,B)) \wedge (C \sqsubseteq A) \wedge (B \sqsubseteq D) \Rightarrow (Dis(C,D))$ | 37.5 | 52.4 |
| 13. | (Every $A$ is a (every $B$ is a $C$)) & (every $B$ is a $C$) ⇒ (every $A$ is a $B$) | | |
| | $(A \sqsubseteq (B \sqsubseteq C)) \wedge (B \sqsubseteq C) \Rightarrow (A \sqsubseteq B)$ | 43.8 | 47.6 |
| 14. | (Everything is a $B$) & (every $A$ is a $B$) ⇒ (nothing is an $A$) | | |
| | $(\top \sqsubseteq B) \wedge A \sqsubseteq B) \Rightarrow (A \sqsubseteq \bot)$ | 62.5 | 66.7 |
| 15. | (Every $A$ is a $B$) & (every non-$B$ is an $A$) ⇒ (nothing is an $A$) | | |
| | $(A \sqsubseteq B) \wedge (\neg B \sqsubseteq A) \Rightarrow (A \sqsubseteq \bot)$ | 87.5 | 85.7 |
| 16. | (Every $A$ is related to at least one $B$ under $R$) & (Everything that something is related to under $R$ is a $C$) ⇒ (Everything that something is related to under $R$ is both a $C$ and a non-$B$) | | |
| | $(A \sqsubseteq \exists R.B) \wedge (Rang(R.,C)) \Rightarrow (Rang(R.,C \sqcap \neg B))$ | 43.8 | 38.1 |
| 17. | (Every $A$ is related to at least one thing in either $B$, $C$, or both, under $R$) & (no $B$ is a $C$) ⇒ (nothing is an $A$) | | |
| | $(A \sqsubseteq \exists R.(B \sqcup C)) \wedge (Dis(B,C)) \Rightarrow (A \sqsubseteq \bot)$ | 62.5 | 85.7 |
| 18. | (Every $A$ is related to at least 1 thing in $B$ under $R$) & (every $A$ is related to at most three things in $B$ under $R$) ⇒ (every $A$ is nothing) | | |
| | $(A \sqsubseteq \geq 1R.B) \wedge (A \sqsubseteq \leq 3R.B) \Rightarrow (A \sqsubseteq \bot)$ | 56.3 | 66.7 |
| 19. | (Every $A$ is related to at least one $B$ under $R$) & (every $A$ is nothing) ⇒ (nothing is a $B$) | | |
| | $(A \sqsubseteq \exists R.B) \wedge (A \sqsubseteq \bot) \Rightarrow (B \sqsubseteq \bot)$ | 62.5 | 66.7 |
| 20. | (Every $A$ is related to at least one $B$ under $R$) & (each thing is related to at most one thing under $R$) ⇒ (nothing is an $A$) | | |
| | $(A \sqsubseteq \geq 1R.B) \wedge (Fun(R.)) \Rightarrow (A \sqsubseteq \bot)$ | 56.3 | 81.0 |

### *3.1.3 Procedures*

All participants were collected in a room. First, the participants were provided with three pages of instructions on the basic meaning of graphs or diagrams, but not on particular rules to solve inference tasks. Second, a pretest to check whether they understood the instructions correctly was conducted; they were asked to choose, from a list of three possibilities, the sentence corresponding to a given graph/diagram, as shown in Fig.8. (the potential difference of familiarities of representations is reduced since both groups received substantial instruc-

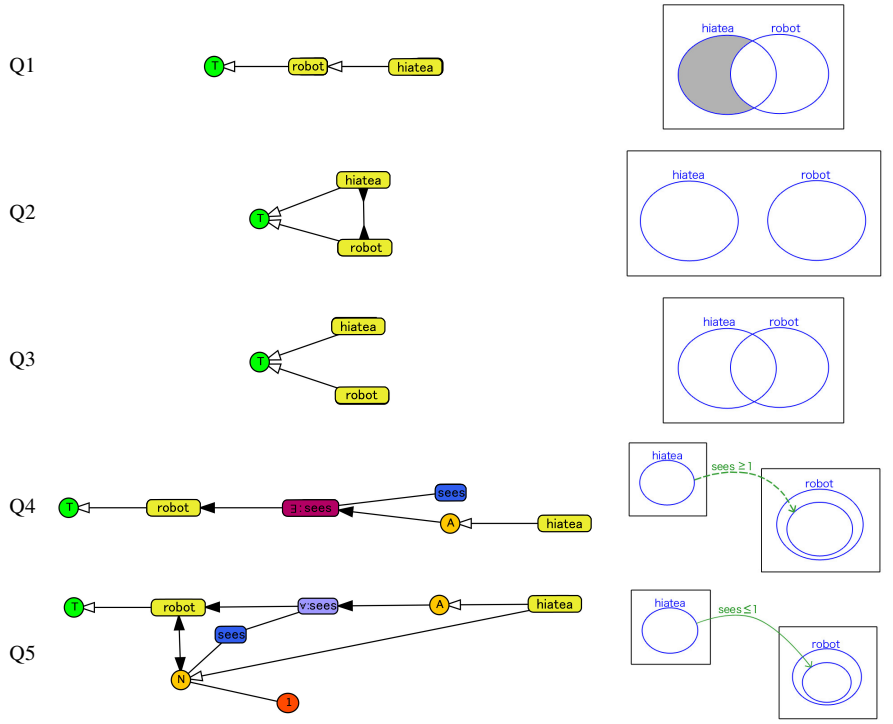


**Fig. 8** Five pretest tasks in topological group (left) and topo-spatial group (right): Answer options in Q1–Q3 are 1. *All hiateas are robots*; 2. *No hiateas are robots*; 3. *None of the above*. Correct answers are 1 for Q1; 2 for Q2; 3 for Q3. Answer options in Q4–Q5 are 1. *All hiateas see at least one robot*; 2. *All hiateas see at most one robot*; 3. *All hiateas see exactly one robot*. Correct answers are 1 for Q4; 2 for Q5.

tion and underwent practice trials: cf. Sato & Mineshima, 2015). Third, the participants were provided with one page of instruction on the meaning of valid transformation (entailment), with two examples of graphs/diagrams: one is valid (corresponding to *All C are B; All B are A*; therefore *All C are A*) and one is not valid (corresponding to *All C are B; All A are B*; therefore *All C are A*). For the full details of the instruction, see the Appendix. After the instruction phase, the participants were asked to solve the main reasoning tasks.

### 3.2 Results and Discussion

The data of the participants who made mistakes in more than two items (out of five) of the pretest were removed. In the following analysis, 3 out of 19 in the topological (SOVA graphs) group, and 4 out of 25 participants in the topo-spatial (concept diagrams) group were removed.

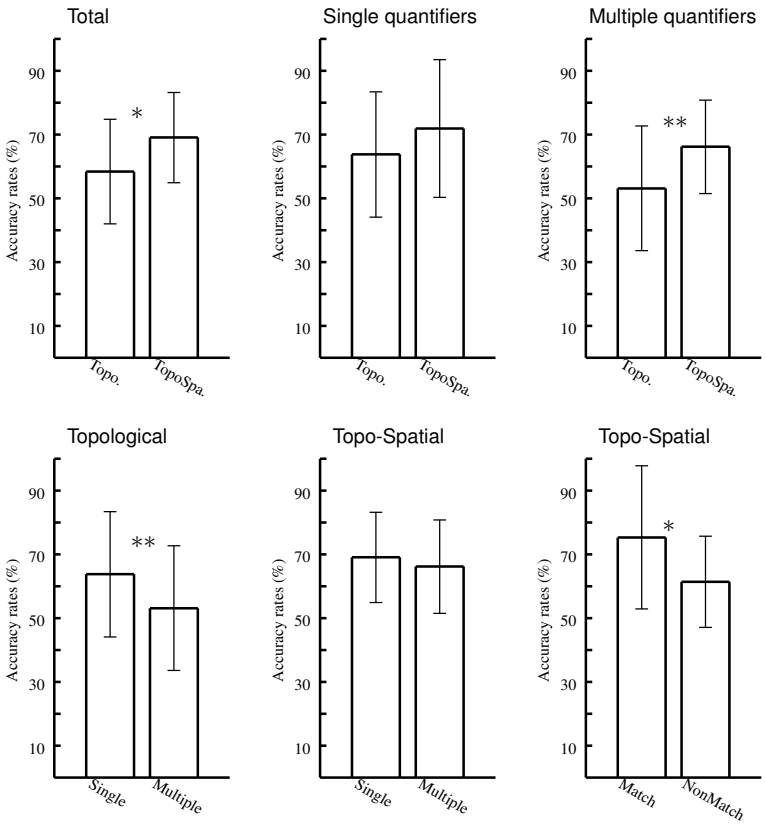


**Fig. 9** Average accuracy rates for inference tasks. Error bars represent standard deviations. $*$ refers to a significant difference at $p < 0.05$.; $**$ refers to $p < 0.01$.

Figure 9 shows the average accuracy rates of inference tasks in the two groups; for each task result, see Table 1. The data was subjected to two-way ANOVA for a mixed design (as is standard, we start by conducting an ANOVA to explore the main effects before we consider our five predictions.). There was a significant main effect of notation factor (i.e., topological versus topo-spatial), $F = 4.435$, $p = 0.042$. There was a significant main effect of factor involving quantifiers (i.e, single versus multtiple quantifiers), $F = 4.712$, $p = 0.037$. There was no significant interaction effect, $F = 0.426$, $p = 0.518$. So each main effect can be considered independently, and the following positive effects were found: effect of topo-spatial representations (69.1%) compared to topological representations (58.4%), as shown in Figure 9 (above, left); effect of single quantifiers (68.4%) compared to multiple quantifiers (60.5%).

**Table 2** Results of ANOVA post-hoc test on the predictions (1)–(4); $**$ refers to $p < 0.01$. *n.s.* refers to no significant difference, $p > 0.05$.

| | | |
|---|---|---|
| -Topo vs. TopoSpa representations in single quantifiers (Prediction 1) | $F = 1.834$, $p = 0.071$ | *n.s.* |
| -Topo vs. TopoSpa representations in multiple quantifiers (Prediction 2) | $F = 2.938$, $p = 0.005$ | $**$ |
| -Single vs. Multiple quantifiers in topological group (Prediction 3) | $F = 2.849$, $p = 0.007$ | $**$ |
| -Single vs. Multiple quantifiers in topo-spatial group (Prediction 4) | $F = 1.532$, $p = 0.134$ | *n.s.* |

A post-hoc test by Bonferroni's method was conducted (see Table 2; the data relevant to prediction 5 is not the target of this ANOVA and a new statistical model is needed.). Regarding the reasoning with single quantifiers, there was no significant difference between the two representations: 63.8% for the topological group and 71.9% for the topo-spatial group where $F = 1.834$, $p = 0.071$. This is consistent with our first hypothesis. Regarding the reasoning with multiple quantifiers, the accuracy rates in the topo-spatial group were significantly higher than those in the topological group: 53.1% for the topological group and 66.2% for the topo-spatial group where $F = 2.938$, $p = 0.005$. This supports our second hypothesis. In the topological group, reasoning with multiple quantifiers was significantly harder than reasoning with single quantifiers where $F = 2.849$, $p = 0.007$. This confirms our third hypothesis. In the topo-spatial group, there was no significant difference between single quantifiers and multiple quantifiers where $F = 1.532$, $p = 0.134$, which is consistent with our fourth hypothesis.

In each comparison between valid items and invalid items, there was no significant difference at the threshold of 5% in two-tailed t-tests. (i) 58.1% for valid items vs. 58.8% for invalid items in the topological group ($t = -0.085$); 71.9 for valid items vs. 66.2% for invalid items in the topo-spatial group ($t = 1.059$). (ii) 66.3% for valid items with single quantifiers vs. 61.3% for invalid items with single quantifiers in the topological group ($t = 0.516$); 79.0% for valid items with single quantifiers vs. 64.8% for invalid items with single quantifiers in the topo-spatial group ($t = 1.878$). (iii) 50.0% for valid items with multiple quantifiers vs. 56.3% for invalid items with multiple quantifiers in the topological group ($t = -0.682$); 64.8% for valid items with multiple quantifiers vs. 67.8% for invalid items with multiple quantifiers in the topo-spatial group ($t = -0.370$). This shows that the accuracy performance is not different between valid and invalid items.

In the topo-spatial group, the averaged data of accuracy rates in the matching cases and in the non-matching cases were subjected to angular transformation and then t-test. The accuracy rates in non-matching cases were significantly lower than those in matching cases: 61.4% for the non-matching cases and 75.3% for the matching cases ($t = 2.748$, $p = 0.012$). This supports our fifth hypothesis.

## 4 General discussion

In summary, our experiment suggests that topo-spatial representations, such as concept diagrams, can be more effective than topological representations, such as SOVA, in reasoning tasks containing multiply quantified information. In topological representations, reasoning

with multiple quantifiers is harder than reasoning with single quantifiers. But in topo-spatial representations, there was no significant difference between these cases. That is, in the topo-spatial case, the difficulty of the inference task did not increase when richer information was conveyed in the premises. In topo-spatial representations, furthermore, reasoning in the cases which require premise diagrams to be merged that contain curves where an inclusion or exclusion relationship is unspecified (non-matching cases) was harder than in cases where the inclusion or exclusion relationships were determined (matching cases).

Regarding the performance difference between single and multiple quantifiers, it is noted that our empirical findings in reasoning with topological representations are similar to those in sentential reasoning, as reported in Nguyen et al. (2012). Our findings shed light on the fact that mental representations elicited from sentences contain simple components only rather than hybrid components (which are an efficient way to represent tasks). Indeed, whether model-like representations based on linked data points (Johnson-Laird et al., 1989; Ragni & Sonntag, 2012; Greene, 1992) or syntactic representations corresponding to parsing trees/graphs of sentences (Braine, 1998), mental representations for expressing multiple quantifiers are assumed to be more complex than those for single quantifiers. Of course, the complexity can be reduced, for example, by using additional representations such as set-theoretical and spatial notions, rather than linking of simple components. However, people not trained in logic and mathematics rarely spontaneously use such additional representations. Accordingly, Bucciarelli and Johnson-Laird (1999) pointed out that the use of Euler circles is a sophisticated method provided by school education, and is not as natural as naive people's mental representations of quantified assertions (p.296).

In the broader context of cognitive science on reasoning, can the result that topo-spatial representations were effective in reasoning with multiple quantifiers provide some more general implications? The findings of effective expression of tasks can contribute to *task naturalisation* in the psychology of reasoning (Politzer, Bosc-Miné & Sander, 2017). If the aim of the current research is measuring people's actual logical (not puzzle solving) capability, tasks involving inference, as opposed to just interpretation, should be set for participants.[9] Our experiment suggests that single quantifiers should be expressed as spatial relations and multiple quantifiers as topological relations, rather than both types being expressed as topological relations. However, according to Politzer et al. (2017), natural tasks of categorical syllogism consist of one premise rather than two premises. This setting is realised by the fact that the existing knowledge of meaningful terms used in the statements can bridge the gap of an unstated premise. This approach in inference research is consistent with the recent developments in natural language processing (Dagan et al., 2009; Bowman et al., 2015) of recognising textual entailment challenge, where entailment relations between two sentences are calculated. Indeed, it is unclear if such a framework using a single premise can be applied to the current case of reasoning about more expressive statements involving quantifiers. However, our findings that there are certain difficulties in merging multiple premises shed light on the question of whether such a process is essential in human (logical) inference. We can explore the question by providing an environment using meaningful content (instead of

[9] For example, errors caused by the misinterpretation of premises should be prevented here, as discussed in the Introduction.

terms with no implied content used in the current study) in which the existing knowledge is available and the task complexity can be controlled.

Furthermore, our findings on inference or entailment judgement are in contrast to those of consistency checking (Sato et al., 2017), where topological representations were more effective than topo-spatial representations. Consistency checking is a kind of logical reasoning in which people are asked to answer the question of whether a diagram makes a contradictory statement. The contrast between these empirical findings suggests that there are two distinct cognitive process underlying logical reasoning from external diagrams. One is a pattern matching strategy, especially based on *syntactic* forms of representations. While patterns of conclusions entailed from premises are unlimited (but in the case of syllogisms, there are some restrictions), there are certain (common) patterns of inconsistency that are exhibited by statements (cf. Flouris et al., 2006). In the tasks of consistency checking, then, the strategy to syntactically match patterns to target representations can be reasonable. This strategy is suitable for the notations which are expressed in a uniform way. Thus, this suggests to us why topological representations were superior to topo-spatial representations in consistency checking tasks. Another strategy is a more *semantic* one, as we now explain. As shown in Fig.5, the conclusion such as *Nothing is an A* requires reasoners to generate some new objects which cannot be found only in syntactic manipulations of representations. Such processes are available only when reasoners correctly understand semantic meanings of given representations. This semantic process might be suitable for the topo-spatial representations.

Our findings also provide a suggestion that efficient systems realising human-like computing, especially interactive theorem proving systems, should be designed in a way to be sensitive to not only syntactic aspects but also semantic aspects. In addition, in the literature of ontology information visualisation, graph (topology) based methods have been dominantly used so far (for a survey, see Ramakrishnan & Vijayan, 2014); by contrast, our findings shed light on the alternative methods using not only topological but also spatial relations. Indeed, our findings suggest that there are merits in using topo-spatial notations for representing ontologies, especially when inferences need to be made.

In this study, we dealt with deductive inference patterns beyond the forms of syllogisms, but, of course, all of them were not covered. For example, proper nouns such as *a is an A* (“a” is an individual constant), as mentioned in the Introduction, were not contained in the corpus of deduction patterns in Nguyen et al. (2012). Although their frequencies might not be so high in the literature of ontology engineering, given the recent findings about fallacies (Mascarenhas & Koralus, 2017), this line of extension would be interesting in our approach using graphical representations. Such an extension should be also applied to the cases using non-standard quantifiers such as proportional quantifiers (e.g., *most*, *few*), which are not definable within first-order logic or OWL 2. In addition, we did not handle relations between verbs, for example, ‘bought’ isInverseOf ‘wasPurchsedBy’, which is a common style of premise in ontology engineering. Some chains of single deduction patterns, as used in Nguyen et al. (2013), are also regarded as an applied type of real world reasoning. So these should be analysed next. Through further extended studies, the nature of human reasoning in general would be explored.

## Appendix: Materials and instructions used in the experiment

https://www.cl.cam.ac.uk/research/ard/exp/MateInst3.zip?attredirects=0&d=1

## Acknowledgements

Parts of this study will be presented in the 40th CogSci Conference (July, 2018). This research was funded by the Leverhulme Trust Research Project Grant (RPG-2016-082) for the project entitled "Accessible Reasoning with Diagrams". The authors would like to thank John Howse, Andrew Blake and Ryo Takemura for their cooperation in the experiments.

## References

Barwise, J., & Shimojima, A. (1995). Surrogate reasoning. *Cognitive Studies: Bulletin of Japanese Cognitive Science Society*, 4, 7–27.

Brachman, R. J., & Schmolze, J. G. (1985). An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9, 171–216.

Braine, M. D. S. (1998). Steps toward a mental-predicate logic. In M.D.S. Braine & D.P.O'Brien (Eds.), *Mental Logic* (pp. 273–331). Mahwah, NJ: Lawrence Erlbaum.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 632–642). The Association for Computational Linguistics.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, 23, 247–303.

Chandrasekaran, B. (2011). When is a bunch of marks on paper a diagram? Diagrams as homomorphic representations. *Semiotica*, 186, 69–87.

Dagan, I., Dolan, B., Magnini, B., & Roth, D. (2009). Recognizing textual entailment: rational, evaluation and approaches. *Natural Language Engineering*, 15(4), i-xvii.

Flouris, G., Huang, Z., Pan, J. Z., Plexousakis, D., & Wache, H. (2006). Inconsistencies, negations and changes in ontologies. In *Proceedings of the 21st National Conference on Artificial Intelligence* (pp. 1295–1300). Menlo Park, CA: The AAAI Press.

Funt, B. V. (1980). Problem-solving with diagrammatic representations. *Artificial Intelligence*, 13, 201–230.

Gentner, D., Holyoak, K. J., & Kokinov, B. N. (Eds.). (2001). *The Analogical Mind: Perspectives from Cognitive Science*. Cambridge, MA: MIT press.

Geurts, B. & van Der Slik, F. (2005). Monotonicity and processing load. *Journal of Semantics*, 22, 97–117.

Glasgow, J., Narayanan, N. H., & Chandrasekaran, B. (Eds.). (1995). *Diagrammatic Reasoning: Cognitive & Computational Perspectives*. Cambridge, MA: AAAI Press/MIT Press.

Greene, S. B. (1992). Multiple explanations for multiply quantified sentences: Are multiple models necessary? *Psychological Review*, 99, 184–187.

Harel, D. (1988). On visual formalisms. *Communications of the ACM*, 31, 514–530.

Hartley, R. T., & Barnden, J. A. (1997). Semantic networks: visualizations of knowledge. *Trends in Cognitive Sciences*, 1, 169–175.

Itzik, N., & Reinhartz-Berger, I. (2014). SOVA - A tool for semantic and ontological variability analysis. In *Proceedings of CAiSE 2014 Forum at the 26th International Conference on Advanced Information Systems Engineering*, CEUR vol 1164 (pp. 177–184).

Johnson-Laird, P. N., Byrne, R. M., & Tabossi, P. (1989). Reasoning by model: The case of multiple quantification. *Psychological Review*, 96, 658–673.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, 138, 427–457.

Khemlani, S., Lotstein, M., & Johnson-Laird, P. (2014). A mental model theory of set membership. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2489–2494). Austin, TX: Cognitive Science Society.

Kroger, J. K., Nystrom, L. E., Cohen, J. D., & Johnson-Laird, P. N. (2008). Distinct neural substrates for deductive and mathematical processing. *Brain Research*, 1243, 86–103.

Mascarenhas, S., & Koralus, P. (2017). Illusory inferences with quantifiers. *Thinking & Reasoning*, 23, 33–48.

Nguyen, T. A. T., Power, R., Piwek, P., & Williams, S. (2012). Measuring the understandability of deduction rules for OWL. In *Proceedings of the 1st International Workshop on Debugging Ontologies and Ontology Mappings, LECP 79* (pp. 1–12), Linköping University Electronic Press.

Nguyen, T. A. T., Power, R., Piwek, P., & Williams, S. (2013). Predicting the understandability of OWL inferences. In *Proceedings of the Extended Semantic Web Conference 2013, LNCS 7882* (pp. 109–123). Heidelberg: Springer.

Politzer, G., Bosc-Miné, C., & Sander, E. (2017). Preadolescents solve natural syllogisms proficiently. *Cognitive Science*, 41, 1031–1061.

Politzer, G., & Mercier, H. (2008). Solving categorical syllogisms with singular premises. *Thinking & Reasoning*, 14, 434–454.

Ragni, M., Singmann, H., & Steinlein, E. M. (2014). Theory comparison for generalized quantifiers. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1222–1227). Austin, TX: Cognitive Science Society.

Ragni, M., & Sonntag, T. (2012). Preferences and illusions in quantified spatial relational reasoning. *Cognitive Processing*, 13, 289–292.

Ramakrishnan, S., & Vijayan, A. (2014). A study on development of cognitive support features in recent ontology visualization tools. *Artificial Intelligence Review*, 41, 595–623.

Sato, Y. & Mineshima, K. (2015). How diagrams can support syllogistic reasoning: an experimental study. *Journal of Logic, Language and Information*, 24, 409–455.

Sato, Y., & Mineshima, K. (2016). Human reasoning with proportional quantifiers and its support by diagrams. In *Proceedings of Diagrams 2016, LNCS 9781* (pp. 123–138). Switzerland: Springer.

Sato, Y., Stapleton, G., Jamnik, M., Shams, Z., & Blake, A. (2017). How Network-based and set-based visualizations aid consistency checking in ontologies. In *Proceedings of the 10th International Symposium on Visual Information Communication and Interaction* (pp. 137–141). New York, NY: ACM.

Sato, Y., Sugimoto, Y., & Ueda, K. (2018). Real objects can impede conditional reasoning but augmented objects do not. *Cognitive Science*, 42, 691–707.

Sato, Y., Wajima, Y., & Ueda, K. (2018). Strategy analysis of non-consequence inference with Euler diagrams. *Journal of Logic, Language and Information*, 27, 61–77.

Schmidt-Schauß, M., & Smolka, G. (1991). Attributive concept descriptions with complements. *Artificial Intelligence*, 48, 1–26.

Steels, L. (1990). Exploiting analogical representations. *Robotics and Autonomous Systems*, 6, 71–88.

Stapleton, G., Compton, M., & Howse, J. (2017). Visualizing OWL 2 using diagrams. In *Proceedings of 2017 IEEE Symposium on Visual Languages and Human-Centric Computing* (pp. 245–253). Los Alamitos, CA: IEEE Computer Society Press.

Stenning, K. (2002). *Seeing Reason: Image and Language in Learning to Think*. Oxford: Oxford University Press.

Stenning, K., & van Lambalgen, M. (2001). Semantics as a foundation for psychology: A case study of Wason's selection task. *Journal of Logic, Language and Information*, 10, 273–317.

Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: logic and implementation. *Cognitive Science*, 19, 97–140.

Sugiyama, K., & Misue, K. (1991). Visualization of structural information: Automatic drawing of compound digraphs. *IEEE Transactions on Systems, Man, and Cybernetics*, 21, 876–892.

W3C OWL Working Group (2012). *OWL 2 Web Ontology Language*. Retrieved, Dec. 2017, from http://www.w3.org/TR/owl2-overview/