Preferred Mental Models in Qualitative Spatial Reasoning: A Cognitive Assessment of Allen's Calculus

Markus Knauff, Reinhold Rauh & Christoph Schlieder
University of Freiburg
Center for Cognitive Science
79098 Freiburg i. Br., FRG
Tel.: ++49-761-203-4944 / -4943 / -4945
knauff|reinhold|cs@cognition.iig.uni-freiburg.de

Abstract

An experiment based on Allen's calculus and its transfer to qualitative spatial reasoning, was conducted. Subjects had to find a conclusion $X\ r_3\ Z$ that was consistent with the given premises $X\ r_1\ Y$ and $Y\ r_2\ Z$. Implications of the obtained results are discussed with respect to the *mental model theory* of spatial inference. The results support the assumption that there are *preferred models* when people solve spatial three-term series problems. Although the subjects performed the task surprisingly well overall, there were significant differences in error rates between some of the tasks. They are discussed with respect to the subprocesses of model construction, model inspection, validation of the answer, and the interaction of these subprocesses.

Introduction

In a spatial three term series problem, two spatial relational terms, $X\ r_1\ Y$ and $Y\ r_2\ Z$ are given as premises. The goal is to find a conclusion $X\ r_3\ Z$ that is consistent with the premises. Such compositions of binary spatial relations have been studied in cognitive psychology (e.g. Johnson-Laird, 1980; Johnson-Laird & Byrne, 1991) as well as in spatial reasoning – a subfield of AI which studies formalisms of encoding spatial relations. One of the most important approaches in this research area is given by Allen's calculus. Allen (1983) presented a temporal logic based on intervals representing events, qualitative relations between these intervals and an algebra for reasoning about relations between these intervals. Although Allen's theory was originally developed in the area of temporal reasoning it has triggered numerous research enterprises in spatial reasoning as well. Güsgen (1989), Mukerjee & Joe (1990), Hernández (1994) and in particular Freksa (1991) transferred Allen's theory to the spatial domain.

Allen denotes thirteen qualitative relations between two intervals: *before (<)* and its converse *after (>)*, *meets (m)* and *met by (mi)*, *overlaps (o)* and *overlapped by (oi)*, *finishes (f)* and *finished by (fi)*, *during (d)* and *contains (di)*, *starts (s)* and *started by (si)*, and *equal (=)* that has no converse. Figure 1 gives pictorial examples for these relations. It is easy to see that these 13 relations can be used to express any qualitative relationship that can be held between two intervals.

In the second part of his paper Allen introduced a reasoning algorithm based on these relations. For instance, if the system receives the information that *X meets Y* and *Y is during Z* it is computed that the following relations between X and Z are possible: *X overlaps Z* or *X is during* Z or *X starts with Z*. The set of all possible conclusions that has $X\ r_1\ Y$ and $Y\ r_2\ Z$ as its premises can be denoted as $c\ (r_1\ r_2)$. Since Allen's theory contains thirteen relations, we get 144 compositions $c\ (r_1\ r_2)$, omitting the trivial "*equal*" relation. It is easy to see that what the system does is very similar to what cognitive scientists call a *spatial three term series problem*.

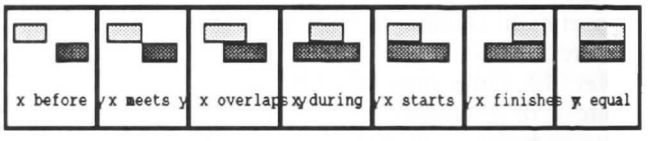

Figure 1: The possible qualitative relations that hold between two intervals (Allen, 1983).

Conceptual and inferential adequacy

Allen's theory has often been claimed to be cognitively adequate (e.g. Freksa, 1992). However, there is no agreement on what cognitive adequacy is. In a strong sense, it at least means that something is a model of human cognition (Strube, 1992). The following is an attempt to give a specialization of what the cognitive adequacy of Allen's theory may be. The question is, whether cognitive adequacy is claimed for the thirteen qualitative relations or for the reasoning mechanism? We distinguish between two kinds of cognitive adequacy, namely *conceptual* and *inferential cognitive adequacy*.

Conceptual cognitive adequacy can be claimed, if and only if empirical evidence supports the assumption that Allen's system of relations is a model of people's conceptual knowledge of spatial relationships. However, as far as we know there is no attempt to test empirically whether or not Allen's calculus is cognitively adequate from this point of view.

Inferential cognitive adequacy can be claimed, if and only if the reasoning mechanism of the calculus is structurally similar to the way of people reason about space. According to this definition, we have to answer two main questions. If the same spatial relational terms, $X\ r_1\ Y$ and $Y\ r_2\ Z$ are given as premises to human subjects and to a system using Allen's calculus, do both come up with the same conclusion? This is the first question, which is relatively easy to test empirically. The second question is concerned with the mental processes underlying spatial inference. For a long time, the great majority of researchers believed in mental representations of formal rules of inference. They proposed that inferences are made by formal derivations and proof of conclusions (Hagert, 1985). Recently, the most successful approaches, however, argued against these rule-based approaches of spatial reasoning. The most important theories are those that go back on the different variants of *mental models* (Huttenlocher, 1968; Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). Generally, the key idea of *mental model theories* is that people translate an external situation of the real world as well as abstract concepts into a mental "simulation" or *mental model*. For spatial inference tasks Johnson-Laird argued that people solve them by imagining the spatial layout. Spatial reasoning from this point of view relies not primarily on syntactic operations, but on the semantic processes of construction and manipulation of mental models (Johnson-Laird & Byrne, 1991). First an internal model of the "*state of affairs*" that the premises describe is built (*comprehension phase*), then a parsimonious description of the mental model including a putative conclusion is constructed (*description phase*). In a third stage, people try to find alternative models of the premises in which this conclusion is false (*validation phase*). If they cannot find such a model, the conclusion must be true. If they find a contradiction, they return to the second stage – and so on until all possible models are tested (Johnson-Laird & Byrne, 1991). Johnson-Laird's theory of spatial inferences does not answer a number of questions, however. In particular, it does not seem plausible that the sequence of tested mental models is a stochastic one. Contrary to that, we believe that there are *preferred models*, which will be generated first. If such *preferred models* for most of the inference tasks described by the Allen calculus can be found, we would assume that – at least in those cases – the model construction process is deterministic. In other words, the given premises uniquely determine the model which is constructed. The agreement of experimental data with mental model theory in this essential point would be a strong argument for the applicability of the theory to spatial inferences with the Allen relations. This is the question of whether Allen's calculus can be called *inferentially adequate*, which we will now investigate.

Experiment

To distinguish between conceptual and inferential aspects of Allen's calculus, the computer-aided experiment was separated into three blocks: a *definition* -, a *learning* -, and the *inference phase*.

Subjects: 33 students of the University of Freiburg, with an age range from 20 to 42 years.

Method and Procedure: In the *definition phase*, subjects read descriptions of the locations of a red and a blue interval using the 13 qualitative relations (in German). Each verbal description was presented with a short commentary about the location of the beginnings and endings of the two intervals and a picture with a red and blue interval that matched the description.

The *learning phase* consisted of blocks of trials, where subjects were presented with the one-sentence description of the red and blue interval. They then had to determine the beginning and ending of the blue interval: out of 8 possible points that were displayed below a red interval, subjects had to choose 2 of them by pressing associated numbers on the keyboard. After confirmation of her/his final choices, the subject was told whether her/his choices were correct or false. If they were false, additional information about all correct answers was given. Trials were presented in blocks of all 13 relations in randomized order. The learning criterion for one relation was accomplished if the subject gave correct answers in 3 consecutive blocks of the corresponding relation. The learning phase stopped as soon as the last remaining relation reached the learning criterion. Subjects needed 15 to 30 minutes to accomplish the learning phase. For each trial, the subject's choices of beginnings and endings, type of answer (correct vs. incorrect) and response times were recorded.

In the *inference phase*, the main part of the experiment, subjects had to solve 144 spatial three–term series problems according to Allen's twelve relations (without "equal"). They were presented in randomized order in the following way: *The red interval starts with the green interval* **(premise 1)**; *The green interval overlaps the blue interval* **(premise 2)**. *Which relationship can be held between the red and blue interval* **(question for conclusion)**? Subjects had to determine the beginning and ending of the blue interval by pressing numbers on the keyboard as learned before. The dependent measure were reaction times and error rates.

Results

For the statistical analyses, a level of significance of 5% will be adopted. Types of answer (correct vs. incorrect) were analyzed, and are reported in the following section.

Learning phase. As mentioned above, the learning phase should guarantee that subjects acquire the relational

Table 1: Mean number of learning trials and percentage of correct answers for each relation.

	<	m	o	fi	di	si	=	s	d	f	oi	mi	>	Total
$\bar{X}$	3.27	3.27	3.36	3.24	3.33	3.42	3.00	3.45	3.55	3.36	3.67	3.67	3.27	3.38
% correct	94.4	92.6	91.9	95.3	93.6	92.0	100	91.2	89.7	93.7	90.1	89.3	93.5	92.9

concepts and associate them correctly with natural language expressions. In Table 1, the mean numbers of learning trials across subjects are listed, that means how many trials our subjects needed to accomplish the required 3 consecutive correct answers for each relation. As can easily be seen, subjects understood the relation "=" at once ($\bar{X} = 3.00$) whereas, on the average, they needed the most learning trials for the relations **"oi"** and **"mi"** ($\bar{X} = 3.67$, each). The maximum number of learning trials that one subject needed for the relation **"si"** was 10. As expected, the pattern was nearly the same for the related measure of the percentage of correct responses (see Table 1). From these results, we can conclude (i) that our learning phase was successful, (ii) that Allen's relations can be acquired and associated with natural language expressions in reasonable time, (iii) and that the following results of the inferential phase are affected only or mostly by the cognitive inferential process. The substantial differences between relations indicate that further research regarding the conceptual adequacy is necessary.

Preferred mental models. To test the global hypothesis that there are generally preferred mental models, a chi-square test was conducted, based on the H_0 that the number of answers per category would be equally distributed for all three term series problems with multiple models (i.e., 72 out of 144 problems; see shaded cells in Table 2). Based on this hypothesis, we obtained a chi-square value of $\chi^2_{(240)} = 1848.04$, $p < .001$. Thus, we can reject the H_0 and adopt the hypothesis that there are preferred mental models in spatial reasoning based on Allen's calculus. Testing the 72 multiple model problems separately, we obtained statistically significant chi-square values in 53 (+ 6 = 59) out of 72 tests (see Table 2 for details)[1]. The most impressive example is the problem **di – oi**, where 84.8% of our subjects chose the relation **"oi"**, whereas the other two correct relations **"di"** and **"si"** (see Table 1) were not used.

Considerable differences can be found with regard to percentages correct, which range from 60.6% (fi – m) up to 97.0% (< – <, m – m, o – <, si – si, d – <) in the one model problems, and from 69.7% (oi – m) up to 100.0% (o – o, o – di, o – d,, fi – >, di – <, di – d, si – <, s – si, > – o, > – d) in the multiple model problems.

1. For the 6 three-term series problems with 9 and 13 models, respectively, expected cell values are so small that the approximation of the test statistic to the chi-square distribution may be rather unsatisfactory. So, these results may be interpreted only with caution.

Discussion

As was outlined in the first section, the question of whether Allen's calculus is a cognitively adequate model for human spatial reasoning has to be broken down into several subordinated questions. To these subordinated questions, we can now give some tentative answers. We could show that (i) Allen's relations together with natural language expressions can be acquired quickly, and (ii) that subjects are rather good in solving inference problems of the given kind (more than 60% correct answers in the worst case). There is also substantial evidence, however, that there are differences with respect to the conceptual and the inferential adequacy, which a cognitive plausible model must account for. With respect to inferential adequacy, an extension and specification of the mental model approach seems most promising to us.

We were primarily interested in the existence of preferred models, but there is also a connection between the error rates found and the mental model approach that is worthy of discussion. Any detailed modeling of the inference process will have to account for the fact that, although the subjects performed surprisingly well overall, there are significant differences in the error rate between some of the tasks. In the following, we will address two questions related to a mental model interpretation of error rates: (1) How do mental model theories explain differences in error rates? (2) Which of these alternative explanations are ruled out by the present data?

To answer the first question it is necessary to come back to the general structure of the inference process described above. First, what Johnson-Laird & Byrne (1991, p.36) call *comprehension phase*, that is, the construction of a model of the premises, second, the *description phase* in which the model is inspected to find which conclusion holds, finally the *validation phase* which consists of examining alternative models. Note that the validation phase amounts to an iteration of the two earlier phases: new models are constructed and inspected until all or a suitable number of models are exhausted. In general, we may expect the three phases to appear in any kind of such inference tasks. However, there are slight differences depending on the inference paradigm that is used. Two inference paradigms are commonly found in literature on reasoning. We will refer to them as *inference verification task* and *active inference task*. To make the difference explicit we introduce the notation $\{\varphi_1,\varphi_2\} \rhd \varphi_3$, to denote the

Table 2: Percentages of correct responses and preferred mental models for Allen's compositions.

	<	m	o	fi	di	si	s	d	f	oi	mi	>
<	<: 97.0%	<: 84.8%	<: 90.9%	<: 84.8%	<: 87.9%	<: 84.8%	<: 78.8%	<, d :‡ 54.5% other 3: 39.4% Σ† 93.9%	o : 39.4% other 4: 57.6% Σ 97.0%	o : 51.5% other 4: 45.5% Σ 97.0%	o : 54.5% other 4: 39.4% Σ 93.9%	= : 45.5% other 12: 54.5% Σ 100.0%
m	<: 93.9%	<: 97.0%	<: 78.8%	<: 75.8%	<: 72.7%	m : 90.9%	m : 87.9%	o : 51.5% other 2: 30.3% Σ 81.8%	o : 48.5% other 2: 39.4% Σ 87.9%	o : 72.7% other 2: 18.2% Σ 90.9%	= : 36.4% other 2: 42.4% Σ† 78.8%	oi : 57.6% other 4: 36.4% Σ 93.9%
o	<: 97.0%	<: 72.7%	<: 54.5% other 2: 45.5% Σ 100.0%	<: 48.5% other 2: 42.4% Σ 90.9%	m : 45.5% other 4: 54.5% Σ 100.0%	o : 39.4% other 2: 42.4% Σ† 81.8%	o : 87.9%	o : 48.5% other 2: 51.5% Σ† 100.0%	o , d :‡ 78.8% other 1: 12.1% Σ† 90.9%	= : 39.4% other 8: 51.5% Σ 90.9%	oi : 66.7% other 2: 9.1% Σ 75.8%	> : 54.5% other 4: 36.4% Σ 90.9%
fi	<: 93.9%	m : 60.6%	o : 87.9%	fi : 93.9%	di : 66.7%	di : 93.9%	o : 87.9%	d : 51.5% other 2: 33.3% Σ 84.8%	= , f :‡ 66.7% other 1: 21.2% Σ† 87.9%	oi : 72.7% other 2: 9.1% Σ 81.8%	oi : 75.8% other 2: 12.1% Σ 87.9%	> : 60.6% other 4: 39.4% Σ 100.0%
di	<: 66.7% other 4: 33.3% Σ 100.0%	o : 81.8% other 2: 0.0% Σ 81.8%	o : 90.9% other 2: 6.1% Σ 97.0%	di : 84.8%	di : 84.8%	di : 72.7%	o : 63.6% other 2: 21.2% Σ 84.8%	= : 48.5% other 8: 51.5% Σ 100.0%	oi : 69.7% other 2: 6.1% Σ 75.8%	oi : 84.8% other 2: 0.0% Σ 84.8%	oi : 72.7% other 2: 3.0% Σ 75.8%	> : 60.6% other 4: 36.4% Σ 97.0%
si	<: 45.5% other 4: 54.5% Σ 100.0%	o : 75.8% other 2: 12.1% Σ 87.9%	o : 75.8% other 2: 12.1% Σ 87.9%	di : 87.9%	di : 66.7%	si : 97.0%	= : 51.5% other 2: 45.5% Σ 97.0%	d : 48.5% other 2: 24.2% Σ 72.7%	oi : 90.9%	oi : 81.8%	mi : 81.8%	> : 93.9%
s	<: 84.8%	<: 75.8%	o : 42.4% other 2: 45.5% Σ† 87.9%	o : 45.5% other 2: 48.5% Σ† 93.9%	fi : 30.3% other 4: 45.5% Σ 75.8%	si : 48.5% other 2: 51.5% Σ† 100.0%	s : 93.9%	d : 81.8%	d : 69.7%	oi : 66.7% other 2: 18.2% Σ 84.8%	mi : 78.8%	> : 84.8%
d	<: 97.0%	<: 81.8%	o : 57.6% other 4: 39.4% Σ 97.0%	o : 57.6% other 4: 30.3% Σ 87.9%	= : 42.4% other 12: 57.6% Σ 100.0%	oi : 48.5% other 4: 39.4% Σ 87.9%	d : 69.7%	d : 87.9%	d : 84.8%	oi : 51.5% other 4: 39.4% Σ 90.9%	> : 66.7%	> : 93.9%
f	<: 84.8%	m : 78.8%	o : 75.8% other 2: 6.1% Σ 81.8%	fi : 42.4% other 2: 54.5% Σ† 97.0%	di , oi :‡ 60.6% other 3: 21.2% Σ 81.8%	oi : 51.5% other 2: 42.4% Σ 93.9%	d : 78.8%	d : 75.8%	f : 93.9%	oi : 42.4% other 2: 45.5% Σ† 87.9%	> : 69.7%	> : 87.9%
oi	<: 51.5% other 4: 42.4% Σ 93.9%	o : 51.5% other 2: 18.2% Σ 69.7%	= : 30.3% other 8: 57.6% Σ 87.9%	oi : 51.5% other 2: 36.4% Σ 87.9%	mi : 42.4% other 4: 54.5% Σ 97.0%	mi : 39.4% other 2: 54.5% Σ† 93.9%	d : 48.5% other 2: 48.5% Σ 97.0%	oi : 51.5% other 2: 42.4% Σ 93.9%	oi : 90.9%	> : 60.6% other 2: 33.3% Σ 93.9%	> : 87.9%	> : 87.9%
mi	<: 39.4% other 4: 54.5% Σ 93.9%	si : 36.4% other 2: 45.5% Σ† 81.8%	oi : 60.6% other 2: 30.3% Σ 90.9%	mi : 93.9%	> : 69.7%	> : 84.8%	oi : 48.5% other 2: 33.3% Σ 81.8%	oi : 57.6% other 2: 15.2% Σ 72.7%	mi : 90.9%	> : 75.8%	> : 90.9%	> : 78.8%
>	= : 39.4% other 12: 60.6% Σ 100.0%	oi : 48.5% other 4: 39.4% Σ 87.9%	oi : 63.6% other 4: 36.4% Σ 100.0%	> : 90.9%	> : 87.9%	> : 87.9%	oi : 42.4% other 4: 54.5% Σ 97.0%	> : 30.3% other 4: 69.7% Σ 100.0%	> : 87.9%	> : 78.8%	> : 84.8%	> : 90.9%

Note. White cells: one model problems; Shaded cells: multiple model problems.

†. Chi-square value n.s.

‡. Both of these relations had the same frequency. Percentages were summed up.

fact that the conclusion φ_3 is compatible with the premises φ_1 and φ_2. Then, the two paradigms may be written as follows:

(1) *inference verification:* does $\{\varphi_1,\varphi_2\} \rhd \varphi_3$ hold?

(2a) *active general inference:* find all φ_3 such that $\{\varphi_1,\varphi_2\} \rhd \varphi_3$.

(2b) *active particular inference:* find some φ_3 such that $\{\varphi_1,\varphi_2\} \rhd \varphi_3$.

The experiment clearly follows (2b), the *active particular inference paradigm*. For active inferences the following structuring of the inference process is implied by mental model theory (cf. the similar description of Johnson-Laird & Byrne, 1991, p.36 for the inference verification paradigm).

repeat
  model $\leftarrow$ Construct a model from φ_1 and φ_2
  answers $\leftarrow$ Inspect the model for φ_3 and add it to answers.
until there are no more models

In the case of an *active particular inference* the loop is executed just once. There are three places where one can look for algorithmic complexity in the procedure, a complexity which should translate into cognitive complexity, and eventually into higher error rates. Obviously, these places correspond to the three phases. The most relevant among them is the validation phase, i.e. the repeat loop in the above procedure, since it has to cope with the complexity that arises from there being different models to examine. Several alternatives (cf. Johnson-Laird, 1983, p.163) can be envisaged for the validation phase. Two of these which are important are:

(1) generation of models within the validation loop
(2) generation of all models before entering the validation loop.

(1) means that the models are generated within the iterative validation loop described above, that is, model construction and model inspection alternate. Opposed to this, in (2) all models are generated before entering the loop, that is, model construction and model inspection processes are separated. In agreement with the overall assumption of sequential flow of control generation before entering the loop is to be thought of as an iterative process in itself. Both alternatives imply that an inference task involving several models is computationally more complex and thus harder than an inference task involving just a single model. Byrne & Johnson-Laird (1989) were able to show for a certain type of spatial inferences in the verification paradigm that the number of models is in fact an adequate complexity measure in the sense that it allows differences in the error rate between tasks to be explained.

In our experimental setting which follows the active particular inference paradigm, an effect of the number of models on the error rate would only be predicted by alternative (2). Since in our paradigm the repeat loop is entered exactly once, there will be computational costs depending on the number of models only in the case that all models are generated at once before entering the loop. We are now in a position to answer the second of the two initial questions: which of the alternative explanations (e.g. 1 or 2) can be ruled out on the basis of the present data? The data of the experiment do not show a positive correlation between the number of models and the error rate; in fact, the correlation is negative: $r = -.4146$, $p < .001$ counting inferences with 13 models (no error possible), $r = -.3737$, $p < .001$ not counting those inferences. So we may discard alternative (2).

Conclusions and future work

We introduced a specialization of the term "*cognitive adequacy of Allen's calculus*", that distinguishes inferential and conceptual aspects. The experiment followed this distinction. The results indicated the existence of preferred models for spatial inferences with the Allen relations. The findings support an account of the inference process following mental model theory. Further evidence will be needed before a detailed modeling of the inference process is possible. A series of ensuing experiments will be concerned with the question of whether the number of models is an adequate complexity measure for reasoning in the Allen calculus – a question which implies a shift from the active particular inference paradigm to the active general inference or the inference verification paradigm. Besides this, we will carry out experiments to test the assumption of the conceptual adequacy of Allen's calculus.

Acknowledgements

This work has been partially supported by the German Ministry for Research and Technology (BMFT) within the joint project FABEL under contract no. 413-4001-01IW104. We are grateful to Gerhard Strube for helpful comments and to Karin Banholzer and Thomas Kuß for carrying out the experiment.

References

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM, 26,* 832-843.

Byrne, R. M. J. & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory and Language, 28, 564-575.*

Freksa, C. (1991). Qualitative spatial reasoning. In D. M. Mark & A. U. Frank (Eds.), *Cognitive and linguistic*

aspects of geographic space. Dordrecht: Kluwer.
Freksa, C. (1992). Temporal reasoning based on semi-intervals. *Artificial Intelligence, 54,* 199-227.
Güsgen, H. W. (1989). *Spatial reasoning based on Allen's temporal logic* (Technical Report ICSI TR-89-049). Berkeley, CA: International Computer Science Institute.
Hagert, G. (1985). Modeling mental models: Experiments in cognitive modeling of spatial reasoning. In T. O'Shea (Ed.), *Advances in artificial intelligence* (pp. 179-188). Amsterdam: North-Holland.
Hernández, D. (1994). *Qualitative representation of spatial knowledge*. New York, NY: Springer.
Huttenlocher, J. (1968). Constructing spatial images: A strategy in reasoning. *Psychological Review, 75,* 268-298.
Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science, 4,* 71-115.
Johnson-Laird, P. N. (1983). *Mental Models. Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.
Johnson-Laird, P. N. & Byrne, R. M. J. (1991). *Deduction*. Hove(UK): Erlbaum.
Mukerjee, A., & Joe, G. (1990). A qualitative model for space. *Proceedings AAAI-90,* 721-727.
Strube, G. (1992). The role of cognitive science in knowledge engineering. In F. Schmalhofer, G. Strube, & T. Wetter (Eds.), *Contemporary knowledge engineering and cognition* (pp. 161-174). Berlin: Springer.