An Analysis of Growth of Knowledge Based on Concepts and Predicates-A Preliminary Study

Meena Kharatmal and Nagarjuna G.

Homi Bhabha Centre for Science Education, TIFR, Mumbai, India

meena@hbcse.tifr.res.in, nagarjun@gnowledge.org

Using the methodology of Refined Concept Mapping, we re-represent the domain of cell biology of secondary and higher secondary levels of textbooks. In this study, we demonstrate that although the number of concept terms increase progressively, the predicate terms achieve constancy depicting a finite set. This indicates that for acquiring expertise in a domain, a finite set of predicate terms is sufficient. In this paper, we illustrate the methodology and discuss the pedagogical implications of the study.

Keywords: Refined concept map, Predicates, Attributes, Knowledge representation, Cell biology

Introduction

Cognitive development studies, in the context of teaching-learning, compare the conceptual structures of novices and experts in terms of coherence, abstractness, parsimony, integration, explicitness, etc. (Brewer & Samarapungavan, 1991; Karmiloff-Smith, 1995; Nagarjuna, 2006). During the growth of knowledge, it is natural to expect an increase in vocabulary: more the knowledge, more the terms. This indicates that new knowledge is introduced through additional vocabulary. However, the clarity of expressions, parsimony, coherence/ integration that we find in expert's knowledge cannot be accounted for, by a mere increase in vocabulary. Therefore, it would be interesting to study the reasons for the above mentioned characteristics of expert's knowledge. This is the problem addressed in this paper.

In the context of science education, language and technical vocabulary are roadblocks. Typically students find the scientific text unfriendly and difficult because of the use of jargonified 'scientific language' that alienates them from the scientific content (Halliday & Martin, 1993). Two contrasting views are put forward: one group suggests that 'jargons' are unnecessary and that the same meaning can be conveyed using everyday language, and the others suggesting that scientific content can be learned only when the language of science is learned. So if there is a problem in the language, then it is due to the matter of the subject itself. Halliday and Martin (1993), claims that the problem is more often due to the focus that is given on the technical terms and not in the grammar. It is important to realize that technical terms cannot be learnt in isolation, but have to be understood as a part of a larger framework. The problem here, according to Halliday and Martin (1993), is due to not making the meaning explicit, not resolving ambiguities, and inappropriate and excessive usage of grammatical metaphors. We shall address the problem of understanding scientific terms in the paper.

By following semantic holism (Quine, 1953), we characterize knowledge as a network, where the meaning of a node (term) arises by virtue of its position in the neighbourhood of the node, rather than from the node itself. The semantic network is a representation of knowledge hierarchically with interconnected nodes and arcs (Quillians, 1967). In a semantic network model, knowledge stored in frames is a network of nodes and relations (Minsky, 1974). Most domains of declarative knowledge is represented in the form of propositions comprising of subject, predicate and object. Graphically it is represented as a node-arc-node model (Novak & Gowin, 1984; Sowa, 1984). This is the basis of the widely used Resource Description Framework (RDF), a standard for network oriented representation recommended by World Wide Web Consortium (W3C)[1]. An outcome of this effort developed into modelling knowledge in terms of ontologies wherein concepts are referred to as classes, which take the place of subject and object in a proposition, whereas predicates are referred to as object properties (relation names) and data properties (attribute names) (OWL)[2]. We shall carry on this legacy for investigating the problem.

One of the specific objectives addressed in this paper is based on Kharatmal and Nagarjuna's (2008) hypotheses that during cognitive development from a novice into an expert:

(i) conceptual change happens due to re-writing the names of relations, and not merely due to re-writing the names of concepts,

(ii) the rate of increase of relation names progressively decrease,

(iii) the same relation names are used consistently eliminating ambiguity and

(iv) the number of relation names required for a formal representation in a given domain is not only finite but few. The lesser the relation names, the greater the formal character of the representation.

We report here some of the empirical results that support the above hypotheses. In the following sections, we describe the methodology, observations, analysis and discussion followed by implications to education and epistemology.

Methodology

The need to focus on the predicate terms for resolving ambiguity and introducing rigor in representation of scientific knowledge was explicated by Kharatmal and Nagarjuna (2006, 2008, 2010). Refined Concept Mapping (RCM) is a methodology that uses a minimal and a least ambiguous set of relation names consistently to represent a body of knowledge. RCM is a development over the traditional concept map (TCM). To cite a few illustrations, the linking words in a textbook sentences, such as–-*is a, can be, has, may be,* etc. are often used in TCM. The problem with these kinds of linking words is that they do not portray the exact meaning and at times, gives rise to ambiguity (Sowa 2006). Our approach has been towards resolving this ambiguity in the representation, by focusing on the usage of the above linking words and replacing them with semantically accurate relation names (linking words) such as–-*part of, includes, surrounded by, located in, has function*, etc. We create concept maps by applying the relation names defined in the relations ontology (RO)[3] of the formal knowledge representation group, the Open Biological and Biomedical Ontologies (OBO)[4], and thus we get what are referred as refined concept maps (RCM) (Kharatmal & Nagarjuna 2006, 2010). Another important criteria to be followed in RCM is to use the relation names consistently throughout the domain. Emphasizing on relation names not only facilitates in disambiguity, but also helps to maintain parsimony in scientific representation. The current research aims at re-representing the scientific knowledge in terms of concept terms and predicate terms, which we refer to, as Refined Concept Mapping (RCM). We follow the Knowledge Representation (KR) model in which a proposition is mapped as [concept name] → (relation name) → [concept name]; [concept name] → (attribute name with value).

For our study, the domain selected is of cell structure and function, in classes 8, 9, 11 textbooks (NCERT 2007), focusing on (limiting to) the topic of cell and its organelles. As this topic is common in 8, 9, 11 classes it could enable us to study the changes in the text when the complexity increases. The basic elements of the study are concept names and predicate terms[5]. The concept names are mostly the scientific terms. For instance, *cell, nucleus, plastids,* etc. are concept names. The predicate terms include relation names and attribute names, and are mostly depicted using the natural language. A few examples of relation names are ---*part of, surrounded by*; and a few attributes are---*has size, has shape*. To begin with, the verbatim text is marked and each verbatim sentence is noted. In these sentences, the linking words are highlighted. These are then replaced with well-defined relation names and attribute names thus transforming them into RCM propositions (see Table 1). Our emphasis is only on the relation names, as part of methodology, we try not to change the concept names.

Since our domain is biology, we draw from the Open Biological and Biomedical Ontologies (OBO) foundry which is collaboratively developing and publishing well-defined relations that are released as the OBO Relation Ontology (RO). For example the definitions of *part of* and *located in* are given below:

> *part_of =def. For continuants: C part_of C' if and only if: given any c that instantiates C at a time t, there is some c' such that c' instantiates C' at time t, and c *part_of* c' at t.*
>
> *located_in =def. C located_in C' if and only if: given any c that instantiates C at a time t, there is some c' such that: c' instantiates C' at time t and c *located_in* c'.*

The rationale for choosing relation names from the RO is that each and every relation name in the relations ontology has a well-defined semantics, thus making the propositions precise and unambiguous. In the light of OBO, the relation names are categorized as *foundational, temporal, spatial, participation* (Smith, et al., 2005). The relation names are chosen based on the classification scheme based on the dimension–*inclusion (class, meronymy, (component-object, member-collection, portion-mass, stuff-object, phase-activity, place-area, feature-event), spatial), possession, attachment, attribution, antonym, synonym, case* (Winston, Chaffin, & Herrman, 1987). In the theory of conceptual representations, the attribute names and their values are represented in the form of the domains–*color, shape,* etc. and region–*red, round*, etc. respectively, of any given object (Gardenfors, 2000).

Now, we shall illustrate this methodology to transform sentences from textbook into RCM propositions by replacing the linking words by well-defined relation and attribute names as shown in Table 1. RCM propositions follow the *subject-predicate-object* structure and it need not be grammatically correct, hence we eliminate the use of articles, prepositions in the RCM.

	Verbatim Sentences	TCM	RCM	RCM Propositions	
1	sharks *can be* great white shark, tiger shark	*can be*	*includes*	sharks *includes* great white shark, tiger shark	1'
2	shark teeth *can be* big, small	*can be*	*has size*	*it is possible that*, shark teeth *has size* big, small	2'
3	nucleus *is a* double layered membrane structure;	*is a*	*enveloped by*	nucleus *enveloped by* double layered membrane structure	3'
4	nucleus *is one of the* organelles in a cell	*is one of the*	*part of kind of*	organelles *part of* cell; nucleus *kind of* organelles	4 '
5	nucleus *is present in* each living cell	*is present in*	*part of*	nucleus *part of* each living cell	5'
6	nucleus *is* small in animal cell	*is*	*part of has size*	nucleus *part of* animal cell; nucleus *has size* small	6'
7	mitochondria *have* DNA and ribosomes	*have*	*consists of*	mitochondria *consists of* DNA and ribosomes	7'
8	mitochondria *have* 2 membrane covering	*have*	*enveloped by*	mitochondria *enveloped by* membrane; membrane *has number* 2	8'
9	plastids *are present* only in plant cells	*are present*	*part of*	plastids *part of* plant cells (only)	9'
10	materials *such as* starch, oils and protein granules	*such as*	*includes*	materials *include* starch, oils, protein granules	10'
11	chloroplasts *are important for* photosynthesis in plants	*are important for*	*has function*	chloroplasts *has function* photosynthesis in s plant	11'
12	plant cells *have* very large vacuoles	*have*	*has size*	vacuoles *has size* large (in plant cells)	12'

Table 1: Verbatim sentences and their conversion into RCM propositions. Notice the *linking words* from the verbatim sentences and their replacement with predicate terms – *relation names* and *attribute names* in the RCM propositions

In 1, 2 we eliminated ambiguous linking word *can be* by replacing it with *includes* in 1' and *has size* in 2'. In the sentences 3-6 one single linking word *is a* is being ambiguously used for four different meanings. This is eliminated by substituting the appropriate relation names *enveloped by, includes, part of, has size* respectively in 3'-6'. The ambiguity of *is-a* link is already being pointed by experts in the field of semantic network (Brachman, 1983; Quillian, 1967). Thus, along with rigor, the substitution also helped in precise expression.

On similar grounds, the linking word *have* used in 7, 8 are portraying two different dimensions, the first one is part and whole dimension and the second one is spatial inclusion dimension. It may implicitly connote the part-whole dimension in the first sentence, but it lacks the precision that is required in the second sentence that of spatial-inclusion. This has been taken care by substituting with *part of* and *enveloped by* in 7', 8'. Similarly, the re-representation has been carried out of the text of 8, 9, 11 classes where structure of cell is discussed (NCERT 2007). The RCM propositions of cell structure and function of classes 8, 9, 11 are graphically represented in the form of refined concept maps and are available at http://gnowledge.org/~meena/cell-biology/

Observations

One of the key points in our research work is of re-representing knowledge with a finite set of predicate terms. We notice that as the depth of the subject levels increase from 8-11, the number of concepts are progressively increasing, but the number of predicate terms do not increase at the same rate as that of concepts, but achieves constancy at a point as depicted in Figure 1. The graph indicates that while concepts increase by a factor of 6, the predicate terms increase by only about 1.5 times.

Constancy in Predicate terms

It may be noted that the number of concepts at class 8, 9, 11 are 75, 195 and 430 respectively. However, at the same three classes, the predicate terms are 10, 15 and 12. The predicates are not increasing at the same rate as the concepts are increasing. We can observe constancy in the number of predicates which also depicts the possibility of a finite set. This denotes that when new concepts get introduced, we do not have to coin a new relation name, but the new concepts can be mapped with a relation name from the given set.

Upon closer observation, we notice that much of the *details* regarding the domain are introduced by using *more attribute names*. We see saturation in relation names but attribute names

do increase at a slower rate, these are 5, 6, 10 in classes 8, 9, 11 respectively (Figure 1). It is interesting to note the proportion of attribute names and relation names in the set of predicate terms. More attribute names appear than the relation names as the level of complexity increases as shown in Figure 2. This indicates that as the depth of the domain increases in its complexity, it is mapped in terms of assigning more of attribute names and less of relation names (please see discussion).

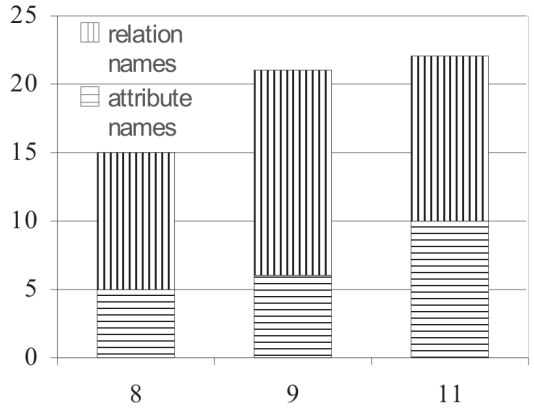

Figure 1: Graph depicting constancy in predicate terms even when the concepts increase progressively in class 8, 9, 11. (Note: the concept names are scaled on secondary y-axis)

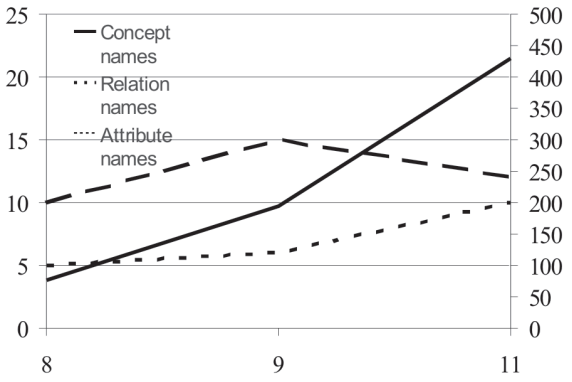

Figure 2: Graph depicting the proportion of relation names and attribute names linked to the concepts in classes 8, 9, 11

To illustrate a specific case, we compare a topic of *mitochondria,* from class 9 and 11, and find that in class 11, there appear 13 new concepts, while only 2 relation names are added which are from the set, and 2 attribute names are introduced. This indicates that when new concept names are introduced, they are linked using a finite and constant set of predicate terms. This explains why the number of predicate terms does not increase as much as the concept names. This observation confirms one of our hypotheses (ii) mentioned earlier.

relation names	C-8	C-9	C-11
includes	15	27	80
consists of	19	36	77
enveloped by	6	12	40
has function	6	45	54
covered by	1	7	11
located in	5	4	24
contains	1	3	43
composed of	3	9	39
discovered	2	2	11
attached to		4	16
called		9	6
produced by		2	
occurs as		3	
appears as		4	
formed by			18
divides by			5
traversed by			2

attribute names	C-8	C-9	C-11
has property	1	6	10
has proportion		1	12
has unit			11
has size	2	1	9
has number	1		9
has shape	1	1	10
has length			4
has color	1	3	2
has arrangement		1	2
has position			2

Table 2: List of predicate terms–relation names and attribute names (emphasized) and the number of concepts that the predicate terms are linked to in class 8, 9, 11, in the chapter on cell biology

Set of predicate terms

As noted above, the predicate terms include the relation names and attribute names. Once the domain is re-represented, it enables us to determine the set of predicate terms that are applied while transforming into RCM. Table 2 shows the list of predicate terms including relation names and attribute names (emphasized) and the number of concepts that these are linked with, in all the three classes. Twenty relation names and 10 attribute names were sufficient to link about 400 concept names applying the RCM methodology. The most widely used predicates are c*onsists of, includes, has function, surrounded by, contains, covered by, located in, has size, has shape, has property, has number.*

Discussion and Implications for Science Education

In this paper, we have discussed the possibility of using a finite set of predicate terms for refining the propositions from secondary and higher secondary school textbook of a domain.

Significance of predicates in Knowledge Representation dates back to Plato. The two widely used relations, type of (subclass of) and instance of (member of), became the basis of formal logic and set theory. These two relations form the core of any inference and they serve as the cement of all thought. The richness of the meaning, however, cannot entirely be captured using only these two relations. This may give us an indication that richness of meaning remains within concept space and not in the predicate space. However, as our study indicates, the required number of predicates do not explode as knowledge increases. The predicates used in science, including those used in Table 2 above for representing biology, are indeed

only those that apply across disciplines because of their general character. They do not remain domain specific, even though concept names are. Thus concept names are explicated by means of a finite set of predicates, irrespective of the domain they may describe. However, we do not rule out the need of special predicates like charge or charm to describe something, say, the fundamental particles.

Predicates are considered the main constituents of symbolic representation based on first order logic (Gardenfors, 2000, p. 37). Modelling of scientific knowledge by focusing on the "predicate space" is elaborated by Gardenfors (2000), who calls them "conceptual spaces". As the recent spurt of activities in semantic web indicate that the knowledge representation studies in computer science also focus on predicate space. Also, they do focus on creating ontologies for different domains. Predicate logic is used explicitly in describing concepts and their relationships, while creating formal ontologies like that of biomedical ontology (Nilsson, 2006). The explicit use of predicates exemplified with constraints in conceptual graphs (Sowa, 1984) has also been one of the insights that we have drawn from focusing on well-defined relation names in characterizing knowledge. A list of well defined set of predicates is available in the appendix of the textbook of conceptual graphs (Sowa, 1984). Selection of attributes for KR has been considered to be of prime importance in designing expert systems. Modelling of real world process requires one to choose a small set of attribute names (Tirri, 1991 cited in Gardenfors, 2000). Also, when we consider the realm of scientific knowledge, the study and changes in the attribute names can represent the dynamics of structure of an object. For representing processes, we often use the change in state space of an object. This change is described by the change in the values of an attribute from prior state to post state. This explains the need to focus on the attribute space.

According to the semantic view of scientific theories, the models are used as predicates in scientific knowledge (Giere, 1992; Stegmuller, 1976; Suppe, 1977; Van Fraassen, 1980). Stegmuller (1976) calls scientific models as complex predicates. Model based reasoning studies by --Nersessian (1999) also suggest that the core of scientific knowledge lies in the modeling of a domain.

Considering that all quantitative predicates (size, volume, position, charge etc.) fall in the attribute space, it is reasonable to assume that the required change in the conceptual structure from a school to a college text consists in explicating the domain in terms of attributes, particularly quantitative attributes. The increase in attribute terms in the current study explains the introduction of the required predicates for building the scientific model as the students progress from school to college level. As a continuation of the study, we intend to explore that the number of predicate terms do not increase substantially from college to university to research levels. Based on our epistemic position, we predict that their number will also remain constant. We do expect similar results for other domains of science, though only further studies will confirm our prediction.

The significance of mapping knowledge using RCM methodology by focusing on the predicate terms has been proposed earlier (Kharatmal & Nagarjuna, 2006, 2010). It was observed in a study that rigor, which is one of the hallmark criteria of scientific knowledge, is rooted in the predicate terms and not in concept names. Further, drawing from this study, a comparison of students, teachers and experts' representations was shown in a subsequent study and it was found that experts tend to focus on the appropriate and consistent usage of the predicate terms (Kharatmal & Nagarjuna, 2008, 2009).

This study continues in the same vein, and further offers new insights and opens up an alternative and simpler way of analyzing the growth of knowledge, specifically scientific knowledge. If this line of thinking is valid, the pedagogical implications could be: the emphasis of training in school and college should be on the use of predicate terms instead on the concept names, for they can be substituted by any variable. The epistemological implications could be: when scientific knowledge becomes increasingly quantitative, the correlations between attribute values should become the focus, when the phenomena are described using functions. Since functions are operationally defined predicates, the meaning is more explicit than the declarative predicates (cognitively grounded predicates) used in early life. Thus during cognitive development, knowledge develops by re-representation of declarative predicates by more and more operationally defined concepts (Nagarjuna, 2006). Thus by focusing on predicate space, it is possible to analyze the growth of knowledge.

In this preliminary study, we attempted to show how a concrete empirical research program to substantiate an epistemological standpoint can be conducted, adding weight to a naturalized evolutionary epistemology program.

Notes

[1] World Wide Web Consortium. http://www.w3.org

[2] OWL: Web Ontology Language. http://www.w3.org/TR/owl-features/

[3] The OBO Relation Ontology. http://www.obofoundry.org/ro/

[4] The Open Biological and Biomedical Ontologies. http://www.obofoundry.org

[5] The usage of the term "linking words" is in the context of TCM or verbatim sentences of the text, while the usage of the term "relation names" and "attribute names" are in the context of RCM propositions.

References

Brachman, R. (1983). What is-a is and isn't: An analysis of taxonomic links in semantic networks. *IEEE Computer*, 16(10), 30-36.

Brewer, W., & Samarapungavan, A. (1991). Children's theories vs. scientific theories: Differences in reasoning or differences in knowledge? In R. R. Hoffman & D. S. Palermo (Eds.), *Cognition and the symbolic processes: Applied and ecological perspectives*, 209-232. New Jersey: Erlbaum.

Gardenfors, P. (2000). *Conceptual spaces–The geometry of thought*. USA: MIT Press.

Geire, R. (1992). *Cognitive models of science*. USA: University of Minnesota Press.

Halliday, M.A.K. & Martin, J. R. (1993). *Writing science: Literacy and discursive power*. London: The Falmer Press.

Karmiloff-Smith, A. (1995). *Beyond modularity: A developmental perspective on cognitive science*. USA: MIT Press.

Kharatmal, M., & Nagarjuna, G. (2006). A proposal to refine concept mapping for effective science learning. In A. J. Canas & J. D. Novak (Eds.), *Concept Maps: Theory, Methodology, Technology*. Proceedings of the Second International Conference on Concept Mapping. San Jose, Costa Rica.

Kharatmal, M., & Nagarjuna, G. (2008). Exploring roots of rigor: A proposal of a methodology for analyzing the conceptual change from a novice to an expert. In A. J. Canas, P. Reiska, M. Ahlberg, & J. D, Novak (Eds.), *Concept Mapping: Connecting Educators*. Proceedings of the Third International Conference on Concept Mapping. Tallinn, Estonia & Helsinki, Finland.

Kharatmal, M., & Nagarjuna, G. (2009). Refined concept maps for science education–A feasibility study. In K. Subramaniam & A. Majumdar (Eds.), *epiSTEME 3 Third International Conference on Review of Science, Technology and Mathematics Education*. Mumbai, India.

Kharatmal, M., & Nagarjuna, G. (2010). Introducing rigor in concept maps. In M. Croitoru, S. Ferre, & D. Lukose (Eds.), *Lecture Notes in Artificial Intelligence: Vol. 6208. International Conference on Conceptual Structures 2010: From Information to Intelligence* (p. 199-202). Berlin, Germany: Springer-Verlag. Doi: 10.1007/978-3-642-14197-3_22

Minsky, M. (1974). *A framework for representing knowledge*. MIT-AI Laboratory Memo 36. http://web.media.mit.edu/~minsky/papers/Frames/frames.html

Nagarjuna, G. (2006). Layers in the fabric of mind: A critical review of cognitive ontogeny. In J. Ramadas & S. Chunawala (Eds.), *Research trends in science, technology and mathematics education, Mumbai*. Homi Bhabha Centre for Science Education.

National Council of Educational Research and Training. (2007). *Science* (textbooks for class 8, 9, 11). New Delhi: NCERT.

Nersessian, N. (1999). Model-based reasoning in conceptual change. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery*, 5-22. NewYork: Kluwer Academic/Plenum Publishers.

Nilsson, J.F. (2006). Ontological constitutions for classes and properties. In H. Scharfe, P. Hitzler, & P. Ohrstrom (Eds.), *Lecture Notes in Artificial Intelligence: Vol. 4068. International Conference in Conceptual Structures 2006,* (p. 37–53). Berlin, Germany: Springer-Verlag.

Novak, J.D. & Gowin, D.B. (1984). *Learning how to learn*. New York: Cambridge University Press.

Quillian, M. (1967). Word concepts: A theory and simulation of some basic semantic capabilities, *Behavioral Science,* 12, 410-430.

Quine, W. (1953). *From a logical point of view. Nine logico-philosophical Essays*. USA: Harvard University Press.

Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A., & Rosse, C. (2005). *Relations in biomedical ontologies. Genome Biology*, 6(5). http://genomebiology.com/2005/6/5/R46

Sowa, J. (1984). *Conceptual structures: Information processing in mind and machine*. USA: Addison-Wesley Publishing Company.

Sowa, J. (2006). Concept mapping. *Talk presented at the AERA Conference*, San Francisco. http://www.jfsowa.com/talks/cmapping.pdf

Stegmuller, W. (1976). *The structure and dynamics of theories*. Berlin, Germany: Springer Verlag.

Suppe, F. (Ed.). (1977). *The structure of scientific theories*. USA: University of Illinois Press.

Tirri, H. (1991). *Implementing expert system rule conditions by neural networks*. New Generation Computing 10, 55-71.

Van Frassen, C. (1980). *The scientific image*. Oxford: Clarendon Press.

Winston, M., Chaffin, R., & Herrman, D. (1987): A taxonomy of part-whole relations. *Cognitive Science*, 11, 417-444.