Design and Development of a Methodology for Relational Content Analysis of Language of Science by Re-representation with Special Reference to Biology

A Thesis

Submitted to the

Tata Institute of Fundamental Research, Mumbai

for the degree of Doctor of Philosophy

in Subject Board of Science Education

by

Meena Kharatmal

Homi Bhabha Centre for Science Education

Tata Institute of Fundamental Research
Mumbai
April 2024

Declaration

This thesis is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgment of collaborative research and discussions.

The work was done under the guidance of Professor Sanjay Chandrasekharan, and Professor Nagarjuna G. at the Tata Institute of Fundamental Research, Mumbai.

Meena Kharatmal

In our capacity as supervisors of the candidate's thesis, we certify that the above statements are true to the best of our knowledge.

Prof. Sanjay Chandrasekharan

Prof. Nagarjuna G.

10 April 2024.

Acknowledgments

Writing the acknowledgments page for a thesis is always challenging. Writing one for a thesis that has taken way longer, a thesis whose academic threads were a lifeline during a literally life-or-death situation is incredibly harder. But let me try. The fact that this thesis exists is in no small measure due to the efforts, support, perseverance and vision of a lot of people.

First and foremost, I am incredibly grateful towards the Homi Bhabha Centre for Science Education (HBCSE) and Tata Institute of Fundamental Research (TIFR), institutions which have always believed in pushing the frontiers of knowledge, and not just allowed but encouraged a Scientific Staff member to pursue a PhD in Science Education. Perhaps next on the list should be wonderful team of doctors and nurses at the BARC Hospital and Jaslok Hospital for their medical care during my quite a few surgeries, that led to complete this work, after a life-threatening condition.

Academically, during my work as a Scientific Staff at HBCSE, I developed a strong interest in mapping knowledge. This interest was shared with the deeply philosophically inclined Prof. Nagarjuna G. who agreed to be my PhD Supervisor. I acknowledge Prof. Nagarjuna for his insights and guidance during my research. I learnt quite a lot from the discussions and interactions with him. I also would like to thank Prof. Sanjay Chandrasekharan who took over the responsibility as the thesis supervisor, after the superannuation of Prof. Nagarjuna. It is always a pleasure to acknowledge Prof. Arvind Kumar (former Centre Director) for his encouragement and interest in my research and for several engaging discussions. I also convey my thanks to Prof. Hemachandra Pradhan (former Centre Director) for supporting my research. I must express my deep gratitude to Prof. Arnab Bhattacharya (current Centre Director) for driving me towards completion of my PhD work, and for personally guiding and supporting me most persistently at the crucial stages of synopsis and thesis completion, seminars, etc. I acknowledge Prof. Arnab for the insight provided on one of my studies with undergraduate students that allowed an extension of my thesis work with international peer acceptance of the ideas. I thank Prof. Arnab for gently but firmly pushing me towards the finish line. I also acknowledge Prof. Shubha Tole, Dean Graduate School, TIFR for providing extension of time as I am also occupied with my regular work as Scientific Staff at HBCSE.

During the inception of the seeds of my research work, I received encouragement and support from several international experts. The visit of Prof. Joel Mintzes, University of North Carolina, Wilmington, USA to HBCSE, and his course on concept mapping and cognitive science was a key driver for embarking me on this research area. I am grateful to Prof. Mintzes for supporting a visiting scholarship at the UNCW. On this trip, I could also visit the Institute for Human and Machine Cognition (IHMC) in Florida, where, I had very encouraging interactions and discussions with several experts. I thank Prof. Pat Hayes, Prof. Robert Hoffman, Prof. Alberto Canas for continuous encouragement. During my participation in various conferences on concept mapping, the epiSTEME series of conferences at HBCSE, conceptual structures conference, systemtic functional linguistics conference, and even email interactions, I had pleasure to interact with experts such as Prof. Joseph Novak, Prof. Ian Kinchin, Prof. John Sowa, Prof. Vinay Chaudhuri, Prof. Yaegan Doran, Prof. Len Unsworth among others. I acknowledge Prof. Richard Barwell for introducing us to the Systemic Functional Linguistics framework.

Participating in the visiting scholar programs and in international conferences has been a memorable highlight of my PhD journey. I thank HBCSE, TIFR; Sir Ratan Tata Trust; Department of Science and Technology; UNCW, USA; IHMC, USA; Mimos Technologies, Malaysia; Technische University, Dresden, Germany; Australian Systemic Functional Linguistics Association, Australia, for their valuable funding support, which enabled these enriching experiences. I also thank the school students and teachers, undergraduate students and college teachers who voluntarily participated in studies of my PhD work.

I enjoyed the academic freedom with flexible working hours at HBCSE (TIFR). The HBCSE Library has been the cozy nook and useful resource of books and journals. At HBCSE, I acknowledge my friends – Sandhya, Smita, Sumana – who have been a moral support during my PhD journey. I acknowledge Sindhu with whom I had engaging discussions. My thanks to Amruta and Yogita too!

I must express my deepest gratitude to my family whose beliefs in my abilities never ended. My family has played an integral role in my accomplishments. I am excited to thank Sunil, Priti, Ravikiran, Sunita, Sachin, Sangita, Kiran and my dearest Arjun, Shreyank and Nirbhay for making my life happy! I am thankful to my father (Bhagwan) and mother (Vijayalaxmi) who pushed the boundaries and continued their support as I was striving to be the first PhD holder in our family. Thank you for everything. I dedicate this PhD to my parents, and hope that I have made them proud and happy. Thanks to my grandparents who always believed in education as a noble profession. And above all, Thanks to the Almighty!

> "Determination is the power that sees us through all our frustrations and obstacles. It helps us in building our willpower which is the very basis of success." - APJ Abdul Kalam.

Dedication

To my family

Abstract

The thesis is about the language of science, specifically biology, from the perspective of relational content analysis of biology textbooks and students' representations. In this exploratory study, we identified gaps and problems in the nature and usage of predicates used for connecting concepts and objects in sentences. To address these gaps, we designed and developed a method of re-representation with iterative refinements focusing on the semantics of predicates, inspired by semantic network-based knowledge representation, concept mapping, representational redescription, and systemic functional linguistics as theoretical frameworks. The method was validated by using well-defined predicates from published open biomedical ontologies and library of generic concepts. The method was used for modelling biology language in terms of using predicates for describing structures, processes, variable properties, experimental procedures, and explanations. This resulted in curating a Reference Set of predicates i.e. relations, attributes. This Set was used for the relational content analysis of cell biology passages from six textbooks as well as students' representations. From this exploratory study, our findings are interesting for unraveling the nature of biology knowledge. We also conducted feasibility, efficacy studies to demonstrate how the method may be used by suggesting a constrained concept mapping method derived from the work. We performed a proximity study to demonstrate how the method can be used as an indicator of expertise. The thesis highlights the importance of developing rigor by reducing ambiguities, and concludes by drawing implications of the work for research in biology education and science communication in general.

Executive Summary

This dissertation focuses on the language of science from the perspective of relational content analysis of biology textbooks and students' representations. In this exploratory work, we have designed and developed a method of re-representation of biology language and show its potential in content analysis of text and students' representations.

The thesis is organized in seven chapters, starting with an introduction and motivation in chapter 1. We summarize some relevant literature, identify gaps and address these gaps by borrowing insights from various theoretical frameworks in chapter 2.

This is followed by chapter 3, where we elaborate on the designing and developing of the re-representation method with illustrations from cell biology topics. In general, sentences are paraphrased in the form - (concept)-[linking-word]-(concept). As the linking words provide meaning, we focus on their kinds and usage to weed out ambiguity, if any, by re-representing them with semantically well-defined predicates, without changing the concepts or loss of meaning. Through re-representing, and iterative refinements, we present a method for modelling biology language comprising of *structures, processes, variable properties, explanations, experimental design*, and creating a Reference Set of predicate vocabulary. The work demonstrates that the cell biology domain can be represented using four broad ontological categories – structure based; class-subclass based; process based; attribution based.

In chapter 4, we present a detailed methodology with coding and categorization of propositions upto 3 levels of the corpus data from school, high school, undergraduate and post graduate levels textbooks and high school and undergraduate students' representations.

We discuss the procedures, data analysis, and findings in chapter 5. In chapter 6, we elaborate on studies about feasibility, efficacy of the method, and proximity percentage as an indicator for developing expertise, all of which are relevant for science education research. In the concluding chapter 7, we discuss the conclusions and implications of the work in biology education and science education in general.

A graphical representation of the work done in the thesis is presented in the Figure 1 below.

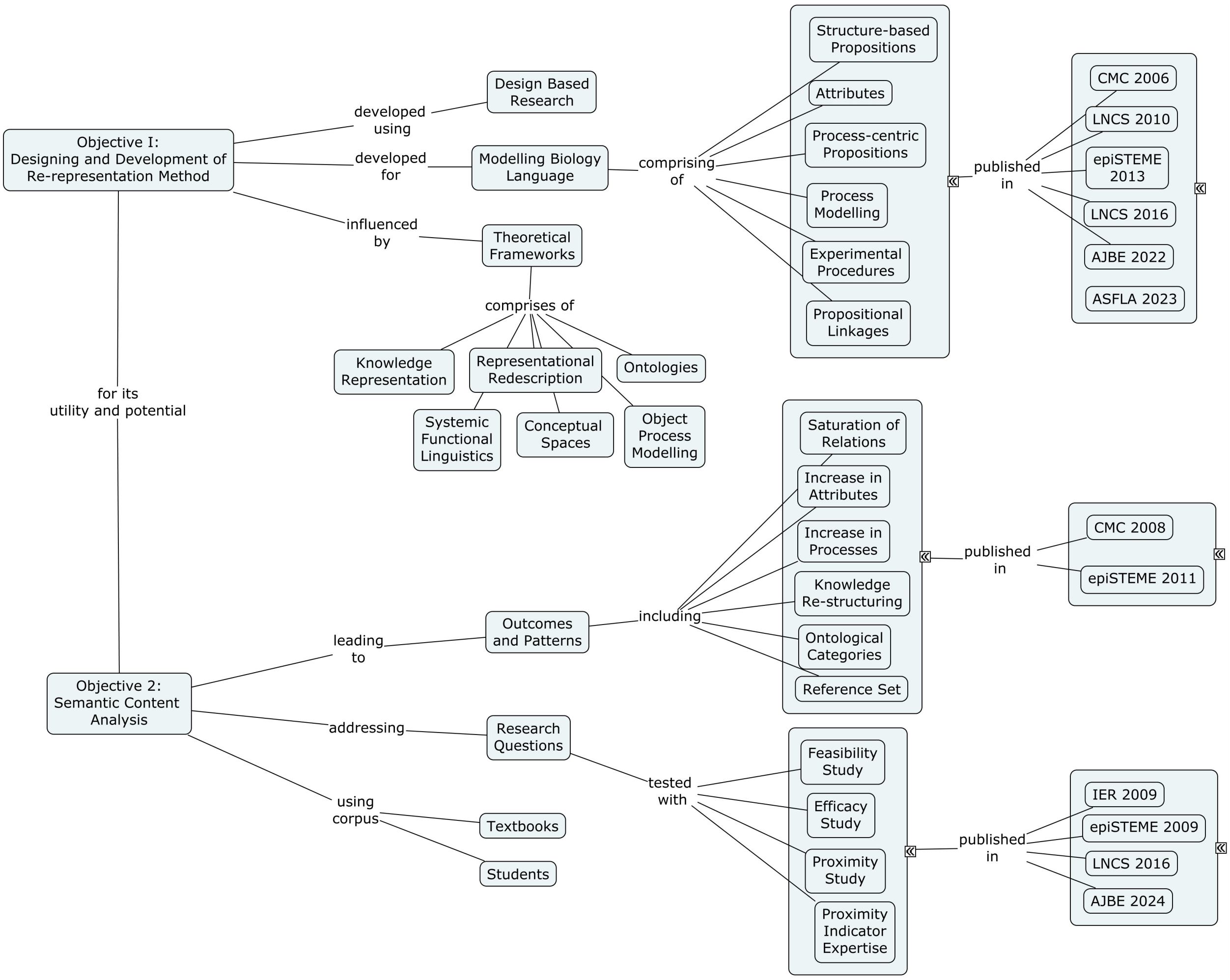


Figure 1: A graphical representation of the research model. (To read from top to bottom and left to right).

Publications

Peer-reviewed Publications

1. Meena Kharatmal & Nagarjuna G. (2024). A Preliminary Study on Effect of Providing Linking Words in Concept Mapping for Representing Cell Biology Knowledge. Asian Journal of Biology Education, (16), p. 9-14. https://doi.org/10.57443/ajbe.16.0_9

2. Meena Kharatmal, Mayur Gaikwad, Aashuthosh Mule, & Jaikishan Advani (2022). Exploring College Biology Students' Understanding of Experiemental Design. Asian Journal of Biology Education, 14(1), p. 2-7. https://doi.org/10.57443/ajbe.14.0_2

3. Meena Kharatmal & Nagarjuna G. (2016). Using Semantic Reference Set of Linking Words for Concept Mapping in Biology. In A. Canas, P. Reiska, & J. Novak (Eds.), Innovating with Concept Mapping. CMC 2016. Communications in Computer and Information Science, vol. 635, p. 315-329. Springer. https://link.springer.com/chapter/10.1007/978-3-319-45501-3_25

4. Meena Kharatmal & Nagarjuna G. (2010): Introducing rigor in concept maps. In M. Croitoru, S. Ferre, and D. Lukose (Eds.), Lecture Notes in Artificial Intelligence. International Conference on Conceptual Structures 2010: From Information to Intelligence, vol. 6208, p. 199-202. Springer-Verlag. https://link.springer.com/chapter/10.1007/978-3-642-14197-3_22

5. Meena Kharatmal (2009): Concept Mapping for Eliciting Students' Understanding

of Science. Indian Educational Review, 45(2), p. 31-43. https://okeanos.files.wordpress.com/2009/11/concept-maps-meena-hbcse-ier-july-2009.pdf

Conference Proceedings

1. Meena Kharatmal (2023a). Field Property for Modeling Biological Process. In Australian Systemic Functional Linguistics Association Conference 2023: Celebrating Semiosis, University of Wollongong, Australia, 21-24 November 2023. https://okeanos.files.wordpress.com/2023/12/asfla-2023-field-property.pdf

2. Meena Kharatmal (2023b). Analysis of undergraduate biology students' multimodal depiction of experimental design using SFL framework. In Australian Systemic Functional Linguistics Association Conference 2023: Celebrating Semiosis, University of Wollongong, Australia, 21-24 November 2023. https://okeanos.files.wordpress.com/2023/12/asfla-2023-multimodal-analysis.pdf

3. Meena Kharatmal (2014). A Proposal to Extend Concept Mapping to Concept Lattices for Representing Biology. International Summer School on Methodology of Task Design – How to Construct Exercises for Learning, Technische University, Dresden, Germany, 2014. https://okeanos.files.wordpress.com/2014/12/summerschool-short-paper.pdf

4. Meena Kharatmal & Nagarjuna G. (2013): Representing Change Using Concept Maps. In G. Nagarjuna et.al. (Eds.) Proceedings of epiSTEME 5 – International Conference to Review Research on Science, Technology and Mathematics Education, p. 124-131. India: Cinnamonteal. https://okeanos.files.wordpress.com/2008/04/representing-change-using-concept-maps-epi5-paper.pdf

5. Meena Kharatmal & Nagarjuna G. (2011): An Analysis of Growth of Knowledge Based on Concepts and Predicates—A Preliminary Study. In S. Chunawala & M. Kharatmal (Eds.) Proceedings of epiSTEME 4 – International Conference to Review Research on Science, Technology and Mathematics Education, p. 144-149. India: Macmillan. https://okeanos.files.wordpress.com/2008/

04/analysis-growth-knowledge-concepts-predicates-epi4-paper.pdf

6. Meena Kharatmal & 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: MacMillan Publishers Ltd. https://okeanos.files.wordpress.com/2008/04/rcm-feasitibility-epi-3-12-meena-gn.pdf

7. Meena Kharatmal & Nagarjuna G. (2008): Exploring the Roots of Rigor: A Proposal of a Methodology for Analyzing the Conceptual Change from a Novice to an Expert. In A. Canas, P. Reiska, M. Ahlberg, J. Novak (Eds.) Concept Mapping: Connecting Educators. Proceedings of the Third International Conference on Concept Mapping. Tallinn, Estonia & Helsinki, Finland. https://okeanos.files.wordpress.com/2008/09/exploring-roots-of-rigor.pdf

8. Meena Kharatmal & Nagarjuna G. (2007): An Alternative Proposal for Eliciting and Assessing Students' Knowledge Structure. In C. Natarajan & B. Choksi (Eds.) epiSTEME 2 Second International Conference on Review of Science, Technology and Mathematics Education. Mumbai: MacMillan Publishers Ltd. https://okeanos.files.wordpress.com/2008/08/episteme-2.pdf

9. Meena Kharatmal & Nagarjuna G. (2006): A Proposal To Refine Concept Mapping for Effective Science Learning. In A. Canas, J. Novak, (Eds.) Concept Maps: Theory, Methodology, Technology. Proceedings of the Second International Conference on Concept Mapping. San Jose, Costa Rica. https://okeanos.files.wordpress.com/2008/08/cmc2006-p151-2.pdf

10. Meena Kharatmal & Nagarjuna G. (2004): Understanding Science Through Knowledge Organizers. In J. Ramadas & S. Chunawala (Eds.) epiSTEME 1 First International Conference on Review of Science, Technology and Mathematics Education. Mumbai: HBCSE. https://okeanos.files.wordpress.com/2008/08/episteme-11.pdf

List of Figures

List of Tables

Contents

Chapter 1

Introduction, Motivation and Objectives

Language of science is a key element in science education. Researchers in the area of philosophy (Carnap, 1938; Woodger, 1938), linguistics and science education (Halliday and Martin, 1993; Lemke, 1990; Kulkarni, 1988; Kharatmal and Nagarjuna, 2008, 2010) have depicted problems in understanding, characterizing, representing and communicating scientific language. The recent areas of computer science, knowledge representation too are focusing on disambiguation, precision in representing knowledge and ontologies (Sowa, 2003; Brachman, 1983; Smith et al., 2005; Keet and Artale, 2008; Bennett et al., 2013).

In science education, conceptual understanding is intertwined with learning the language of science. Therefore it may be affected if the language of learning is not supported (Zukswert et al., 2019). This thesis highlights the representation of scientific language in science textbooks focusing on semantics.

In this chapter, we provide an introduction with background and motivation that led to the idea of a thesis in the area of the language of science in science education.

1.1 Background and Motivation

It is well known that the language of science (academic language) tends to use more controlled vocabulary as compared to everyday language. While everyday language can tolerate ambiguous use of words, contradictions, indeterminacies, etc. (Halliday and Martin, 1993), scientific language cannot. Avoiding misunderstandings, controversies, and errors in analysis cannot be achieved without weeding out ambiguity (Luszczewska-Romahnowa, 1979).

To consider language difficulties in science, we need to address it from the perspective of the nature of science and scientific language, along with students and teachers (Gardner, 1974). The nature of science includes processes, classification, correlation, explanation, etc., which generates new terms that become part of the technical vocabulary. Science has a specialist vocabulary that has not only unfamiliar and new terms but also familiar terms. Even with the familiar terms (energy, work, force), students have to construct new meanings in the context of science (Evagorou and Osborne, 2010). When new terms are introduced in the science vocabulary, such as electric current, these are done employing *metaphorical mechanisms*, for example, depicting the flow of water. These mechanisms serve to introduce new vocabulary into the prevailing language structure, as explicated by Rom Harre (Adúriz-Bravo et al., 2021).

Even though the scientific discourse uses everyday language, there are subtle differences, such as longer sentences and the use of passive voice to describe processes. The jargonized vocabulary and logico-grammatical terms may also not be available for students' cognitive structure. The logico-grammatical terms are used while linking logical sequences of ideas, depicting causality in conditional statements, depicting abstraction, generating hypotheses, etc. The scientific discourse often uses logico-grammatical terms that may interfere with comprehending scientific language. In this context, Lemke (1990) equates learning science as talking science. He further states:

> "Talking science" means observing, describing, comparing, classifying, analyzing, discussing, hypothesizing, theorizing, questioning, challenging, arguing, designing

> experiments, following procedures, judging, evaluating, deciding, concluding, generalizing, reporting, writing, lecturing, teaching in and through the language of science. Lemke (1990, p. 1)

Halliday and Martin (1993) argue that the problems faced while learning science can be dealt with by making the grammatical construction of sentences used in the language of science explicit while teaching. Lemke (1990) also complains that teachers leave the semantics and grammar of scientific language completely implicit for both first-language and second-language learners. The jargonized scientific language also becomes a barrier to conceptual understanding and learning scientific skills (Halliday and Martin, 1993; Zukswert et al., 2019). This often results in excessive focus on the abstract vocabulary of science, mostly the concepts (class terms). Therefore, our emphasis is on the semantic relations based on the insight:

> "Concepts are ideational kernels that in isolation are devoid of meaning. Concepts can only be defined by their relationship to each other. Meaning is a function of the interrelationships between concepts in the knowledge base, and not the concepts per se" (Palmquist and Dale, 1997).

A similar idea is used by Lemke (1990) as a linguist trying to apply this insight in science education. Semantic relations, as these are domain-independent, can become the crucial bridge between everyday language and the language of science, even in science education. Lemke (1990) suggested that teachers must think about semantics explicitly while teaching, and students need to understand thematic patterns based on semantic relations leading toward scientifically correct statements. Further, Lemke (1990) argues that semantic relations are more general and abstract than grammatical relations, though in many ways, they have closer affinities. In a graphical representation, the labels on the lines that link the technical terms, or thematic items, refer to semantic relations (Lemke, 1990; Novak and Gowin, 1984; Fisher, 1990). It is the 'links' i.e., semantic relations, that we are interested in focusing on for representation of scientific language used in the textbook by introducing a method of re-representation.

Representing scientific language is about creating a pattern of relationships of meanings woven with the semantic resources of language. This organic or holistic theory of meaning can be called the network theory of meaning. Semantic relations have been of broader interest among researchers in the area of philosophy (Schwarz and Smith, 2013; Smith et al., 2005), cognitive science (Winston et al., 1987), linguistics (Lemke, 1990), computer science (Quillian, 1968; Sowa, 1984), natural language processing, mainly controlled natural language (Fuchs, 2018; Kuhn, 2010) and ontological engineering (Smith et al., 2005; Ong et al., 2017; Hoehndorf et al., 2015). Knowledge can be represented as a network (Quillian, 1968) or a conceptual graph (Sowa, 1984), with relations as edges linking the various class terms as nodes as concept maps (Novak, 2010; Mintzes et al., 1997; Cañas and Novak, 2008), as SemNet (Fisher and Kibby, 1996) obtaining a network of statements. Eventually, this graph-based representation impacted the data exchange model in the form of RDF, triple stores, and the other compatible graph databases (Hayes, 2004; Ontotext, 1999). As they become part of the formal knowledge representation programs, the semantic relations and logical connectives appear as primitives, vindicating that they form part of a layer beneath class terms, where the latter are introduced as variables.

For modeling language of science, focusing on semantic relations among the class terms, identifying ambiguities is essential (Lemke, 1990). The language of science is organized in hierarchies of kinds and parts (Halliday and Martin, 1993). Therefore, teachers and students need to be familiar with some basic kinds of semantic relations, such as "class subclass", "part whole", etc., so that they can explain meaningful relations among scientific terms unambiguously (Lemke, 1990).

Scientific vocabulary mainly consists of technical terms, i.e., class terms. Given that scientific terms (class terms) are defined by specifying the relations between class terms using linkages, it is reasonable to think that meaning resides in the linkages. By referring to this as the *network theory of meaning*, we base the thesis work on this theoretical ground, focusing on the predicate terms in contrast to the class terms. In this thesis, we apply the term 'predicate terms' for linking words (as depicted in the concept mapping work), similar to object properties and data properties (as depicted in the biomedical ontology work). For example, predicate terms refer to 'is a', 'part of,' 'composed of,' 'present in,' 'made of,' 'occur in,' 'function of,' 'proportional to,' 'has size', 'has color', 'has shape' etc. These are mainly domain-independent and are part of not only the language of science but also everyday vocabulary. Therefore, it is intuitive to pass them tacitly while dealing with class terms explicitly. Most content analysis methodologies, therefore, ignore them, except the relational content analysis (Carley, 1990). In this thesis, we follow the theoretical perspective of network theory of meaning (Churchland, 1988; Quine, 1948) by adopting the methodology of relational content analysis to demonstrate this potential. We take additional inputs borrowed from the growing availability of published ontologies (OBOFoundry, 2020).

The language of science adopted specific grammatical preferences determined due to semantic requirements (Lemke, 1990). Syntax, though is mainly conventional, which may vary not only from one language to another but also within a language. Its invariant logical structure captured in the grammar is not independent of semantics. Analysis of sentences into the subject and verb phrases in linguistics or a statement/proposition into subject, predicate, and object in logic and knowledge representation schemes are well known. The ontology engineers designing semantic web platforms apply the form of graph databases, which are primarily triple stores. The common concern in most of these approaches, by linguists, philosophers, logicians, or ontologists, is to bring the implicit semantics to the foreground by making them explicit.

A significant outcome of these approaches is, that essential semantic relations have been identified and curated by both linguists (Halliday, 1985; Lemke, 1990), and researchers in the area of cognitive science (Winston et al., 1987) in knowledge representation (Markowitz et al., 1992a; Sowa, 1984) and ontology engineering (Smith et al., 2005). Semantic relations have been applied for creating intelligent texts, querying, answering problems, etc. (Barker et al., 2004). However, its application in science education and content analysis of scientific text in the language of science is not fully explored.

Regular teaching and learning focus mainly on developing a vocabulary of scientific concepts through their definitions presented as statements. As a result, the focus is mainly on the vocabulary of concepts, ignoring the vocabulary of predicates (relations and attributes). The predicate terms are taken for granted and, therefore, remain implicit. Our work is an attempt to uncover this implicit into explicit knowledge. In our early research, we observed and reported the ambiguous use of predicates in the textbooks and published concept maps (Kharatmal and Nagarjuna, 2006b). Over the course of our research, we also reported saturation of the number of relations, indicating a limited number of well-defined semantic relations are sufficient in a language of science, even in biology (Kharatmal and Nagarjuna, 2011). We conducted and reported extensive content analysis of passages from biology textbooks (Kharatmal and Nagarjuna, 2016).

The focus on predicates led us to explore whether the consistent use of well-defined linking words (predicates) is a good indicator of rigor in scientific language. This motivated us to develop an iteration-based re-representation method based on well-defined predicates following the widely evolving field of the semantic web, which made the availability of published open biomedical ontologies (repository of predicates and concepts). This research work is formulated with an underlying assumption that the meaning of any term originates from relations between other terms and not concepts per se.

The thesis applies ontology research with an explicit focus on semantics and drawing from the published ontologies (Ong et al., 2017; Smith et al., 2005; Xiang and Janga, 2015) for relational content analysis (Carley, 1990) of science text and is supported by the philosophical insight of network theory of meaning (Churchland, 1988; Quine, 1948). This thesis documents the iterative development of the method and demonstrates its use for the semantic content analysis of biology texts, students' representations, and in modeling the language of biology. A note on the usage of the terms that we use based on the domain of work: Sentences are paraphrased into subject-predicate-object representing a proposition. We will follow a similar syntax considered as [concept-linkingwords-concepts] and [node-link-node] in the concept mapping domain. When applied in semantic web standards, this form is represented as triples of [class-objectproperty-class], and [class-dataproperty-class]. In the logic and knowledge representation language, predicates include relations and attributes, which also correspond to object property and data property respectively in the open biomedical repository. We will follow the usage of concepts for technical terms that are nouns, the usage of relations for verbs/linking words, the usage of attributes for variable properties/adjectives. We often underline modalities and logical connectives that define the scope of each statement and express reasoning, forming complex sentence structures.

1.2 Early Explorations

In this section, we trace the journey of our research work. Initially we started the research due to our interest in the area of concept mapping – a method of graphical representation of a domain (Novak and Gowin, 1984; Mintzes et al., 1997), based on Ausubel et al. (1978) classroom learning theory. The widely recognized facet from the theory is that:

> ... "If I had to reduce all educational psychology to just one principle, I would say this: The most important single factor influencing learning is what the learner already knows. Ascertain this and teach accordingly".

This focuses on the learner's prior concepts in the learner's cognitive structure. It emphasizes the importance of existing knowledge and to build new knowledge in accordance with the *subsumption* theory of classroom learning (Ausubel et al., 1978). The widely used concept mapping method is developed on the Ausubel et al. (1978) learning principles with a graphical format of concept-linkingword-concept, similar to the node-link-node form commonly used in the Knowledge Representation (KR) area. Our interest developed in studying the knowledge organizers in biology (Kharatmal, 2009; Kharatmal and Nagarjuna, 2007, 2006a).

The classical KR tools, such as conceptual graphs (Sowa, 1984), CYC (Lenat and Guha, 1990), knowledge graphs, Word Net, (Fellbaum, 1998; Miller, 1995) are widely used in computer science and knowledge engineering research and development. However, these are not as accessible to researchers in the education community due to their complex design and requirement of technical background to use and understand. On the other hand, the most accessible and easy to use in this genre are mind mapping (Bono, 1977) and concept mapping (Novak and Gowin, 1984). In mind mapping, we think only about the terms or vocabulary. On the other hand, concept mapping leads us to think in terms of named relations, verbs or predicates. As our research is focused on the predicate space, we are using the model of concept mapping. While holding on to this concept mapping model, we will be extending the model by introducing additional conventions to the representation. The concept mapping, when scored, can also be used for content analysis in terms of the number of non-redundant concepts, propositions, hierarchy levels, cross-links, branches, examples, etc. (Novak and Gowin, 1984). These characteristics of the map are being used in the concept mapping community for evaluating the maps and also very often to use them as standard conventions of concept mapping. For example, though every map does not need to have cross-links, the advocates of the method encourage to look for cross-linkages. Therefore, concept mapping already has the potential to make implicit knowledge explicit and also visualize knowledge graphically. Often, studies have reported that the method also helped in gaining new insights, which was not possible when knowledge is represented in a linear text format (Mintzes et al., 1998; Kinchin, 2000).

The concept mapping research has been successful for several decades in various areas of science education research – meaningful learning, misconceptions research, knowledge structure, conceptual change, knowledge organization in biology, etc. (Novak and Gowin, 1984; Novak, 1990; Mintzes et al., 1997; Wandersee, 2002). Further, the tool also inspired to develop the semantic network in knowledge organization of biology (Fisher, 1990; Fisher et al., 2000; Fisher, 2002; Fisher and Kibby, 1996). The semantic network tool certainly had the concept mapping method as its base but its main focus on creating a network with some emphasis on linking phrases. The concept maps were mostly created using paper and pencil, while the semantic networks could be created using computers. Further, a freeware for creating concept maps called CmapTool, is developed (IHMC, 2011; Cañas and Novak, 2008). Since 2004, the concept mapping community has been organizing biennial international conferences highlighting the areas of education, semantics, assessment, etc.

The focus in concept mapping has been on the concepts (class terms, nodes), with little emphasis on the nature of linking words. The method gives freedom in choice of linking words which are from the everyday language vocabulary. For example, during our review of concept mapping work, we noticed that the widely used linking words were: "is", "are", "can be", "has", "have"; and moreover the same linking words were used for different meanings. This results in ambiguity in concept maps (Kharatmal and Nagarjuna, 2006b). As the concept mapping research is based on the meaningful learning principle, then we think that it can contradict the purpose.

To highlight this issue, we conducted a preliminary exploratory study on the critique of the concept mapping method and published it to bring this to the notice of the concept mapping community (Kharatmal and Nagarjuna, 2006b). We discussed the principles of disambiguation, parsimony, and rigor in the critique paper with examples from published concept maps. These principles serve as our guiding principles that are applied to a given knowledge structure. The assumptions with which we began the exploration were based on the nature, structure, and dynamics of knowledge.

We illustrated, with examples, the problems related to (i) the accuracy of linking words and (ii) validating hierarchies. While depicting the problem of accuracy, we pointed out the problems with the usage of 'have,' and 'can be' to mean at least two different meanings – 'part whole' and 'class subclass'. We think this was a serious observation that gave rise to map ambiguity in concept maps. We proposed the semantically well-defined relations: to use 'part of' to mean part whole relations and 'includes' to mean class subclass relations (Winston et al., 1987). We called this as *refined concept maps* (Kharatmal and Nagarjuna, 2006b). In addition, we pointed out the problems in scoring the hierarchy levels in the concept map. It applied the graphical criteria irrespective of ascertaining the use of the same linking words to mean the same expression in the same hierarchy (Mayr, 1969).

The work was noticed in the concept mapping conference and the science education research communities. However, it also received some criticism that restricting the linking words would create obstacles during concept mapping, and not allow the freedom to choose any linking words while mapping. In order to provide a rebuttal to this criticism, we explored another study to demonstrate the feasibility of using a constrained set of linking words during concept mapping. In this study, we selected three homogeneous groups of classroom students and conducted the same task with three different methods: (i)group 1 wrote the passages; (ii) group 2 created concept maps (seeds concepts provided); (iii) group 3 created refined concept maps (seed concepts and seed linking words provided). We showed statistically that in group 3 students were able to depict more accurate propositions as compared to groups 2 and group 1 even with using constrained set of linking words (Kharatmal and Nagarjuna, 2009). This study used 7 kinds of linking words as a constraint set that could be used to create any concept maps with or without seed concepts.

The constrained set of linking words was influenced by the seminal work on taxonomy of semantic relations published in the cognitive science journal by Winston et al. (1987). This work provided the strong base as the relations were not only organized but also well-defined to preserve the semantics. These works culminated on our brainstorming of research program to focusing on the nature and kinds of linking words.

We asked, would this constrained set of just 7 semantic relations alone be enough or sufficient for representing cell structure and function at school level? We embarked on this journey to find out by representing higher levels of textbooks, with the hypothesis that even if the class terms increased the relations would not increase at the same rate. This was demonstrated in the preliminary study conducted with three levels of textbooks (Kharatmal and Nagarjuna, 2011).

With the constrained set of linking words, we could represent an entire cell structure and function chapter from a class 8 textbook. The feasibility study, and the textbook representation (Kharatmal and Nagarjuna, 2008) gave us confidence to focus on the nature of linking words. We wanted to embark on a new program, just as there is a program for creating a nuclear vocabulary (Stubbs, 1986) for making the English language simpler. Therefore with this interest, we asked: Can we explore a research program for creating "Nuclear Predicate" or a "Reference Set" for their use in the language of science?

1.3 Objectives of the Thesis

The Thesis has two objectives:

- Developing a Re-representation Method: wherein we design and develop the method focusing on the predicates in a proposition as a unit in a concept map.
- Using the Re-representation Method for Semantic Content Analysis: to find out if it can bear any significance in understanding the language of science.

1.4 Significance of the Thesis

We believe that the development of the Re-representation Method is the contribution of this thesis, not only for representing scientific knowledge but also for conducting semantic analysis of scientific text or language.

1.5 Outline of the Thesis

The chapter-wise outline of the thesis is as follows:

Chapter 1: Introduction, Motivation, and Objectives. The motivations for the thesis are driven by the semantic network (meaning is in the network) principles, the language of science, educational psychology, explicit knowledge, and semantic principles.

We propose that the e-representation method leading to a Reference Set of well-defined predicates that can act as an interface for facilitating implicit to explicit knowledge.

Chapter 2: Theoretical Frameworks and Literature Review. The thesis applies theoretical frameworks from multiple dimensions for designing and developing the re-representation method. These are (1) knowledge as a network principle for semantic holism, (2) the Representational Redescription model for implicit to explicit knowledge, (3) Nominalization, (4) Open Biomedical Ontologies, and (5) Conceptual

Spaces.

This chapter presents a literature review and discussion about the insights borrowed from these frameworks. We also review research works that focus on linking words, which resulted in identifying gaps. This led to framing the research questions.

Chapter 3: Design and Development of Re-representation Method. The method developed as part of the thesis has the following features for representing knowledge:

1. Re-writing the propositions using well-defined vocabulary of semantic relations from the community-based, validated repository (OBOFoundry, 2020).

2. Making a distinction between relations and attributes ie, variable properties (Ong et al., 2017; Smith et al., 2005; Gardenfors, 2000).

3. Representing process-centric propositions by nominalization (Halliday, 2006).

4. Modelling processes (scientific explanations) as state-change using variable properties (von Wright, 1963; Sowa, 2003; Dori, 2002; Doran and Martin, 2021).

5. Modeling experiments as moment-activities and variable properties (Doran and Martin, 2021; Unsworth, 2020; Halliday and Martin, 1993).

6. Propositional linkages for reasoning using logical connectives (Copi et al., 1972).

   This chapter illustrates Re-representation Rules used in semantic content analysis of scientific text.

Chapter 4: Semantic Content Analysis Using the Re-representation. In this chapter, the re-representation method is explored to find its significance in the semantic analysis of text. In this chapter, we present the analysis of language used in textbooks and students following the relational content analysis method. We present the procedure of coding and categorization of data with illustrations.

Chapter 5: Semantic Content Analysis and Findings. This chapter presents a detailed analysis of biology texts as a corpus leading to the patterns about saturation of predicates, growth of knowledge, etc.

Chapter 6: Studies Using Re-representation. We discuss high school students' findings about feasibility and efficacy studies using a constraint set of predicates. We also discuss the proximity percentage as an indicator of expertise.

Chapter 7: Conclusions and Implications. The thesis concludes with contributions of the study, implications, and future scope of the work.

Chapter 2

Theoretical Frameworks and Literature Review

2.1 Introduction

The theoretical frameworks in the thesis bears an interdisciplinary aptitude, as our work is influenced not only by philosophy and logic but also by linguistics, cognitive science, knowledge representation and science education. The philosophical insights of knowledge as a network is crucial in driving our work focusing on semantics. The cognitive model of Representational Redescription serves as a model for the transformation of implicit representations to explicit knowledge. We are informed by the nominalization technique used in linguistics, drawing on published biomedical ontologies, inspired by modeling conceptual spaces while developing the re-representation method for semantic analysis of language of science.

2.2 Knowledge as a Network

Semantic holism (or meaning holism) attributes the meaning of a word to its relations with other words in an expression (Quine, 1948). In this context, we characterize knowledge as a network where the meaning of a node (term, class) is acquired by virtue of its position with the neighborhood of the node rather than the node itself. Knowledge as a network representation can be traced back to modeling knowledge as frames (Minsky, 1974) and as a semantic network (Woods, 1988). In the semantic network model, knowledge stored in frames (Minsky, 1974) is represented as a network of nodes and relations (Quillian, 1968).

The traditional meaning of semantics is "the scientific study of the relations between signs or symbols and what they denote or mean". Understanding the *semantics* in the semantic network entails the *relationship* between form and meaning rather than either the form or the meaning itself (Woods, 1988). Representing the semantics in the scientific language is through the semantic relations connecting the nodes in statements. For philosophers, the interest is about truth and falsity in propositions. For linguists, the interest is in the disambiguation of sentences as the most important requirement for notation in the semantic specification of natural language. The research groups of computational linguistics and artificial intelligence go beyond the form and meaning of relations to extend to information retrieval and inferencing of a system.

In the interest of science education, in this thesis, these insights are applied for representing scientific language in science text. For example, concept mapping with concepts and linking words are considered for the representation of a domain. Concept mapping is an established method to create a knowledge network with concepts and linking words, creating propositions (Canas et al., 2015; Mintzes et al., 2000; Novak, 2010). Concept mapping is a two-dimensional graphical representation of concepts connected with linking words organized in a hierarchy with branches and cross-linkages (Novak and Gowin, 1984). Research shows that knowledge organization occurs through creating a network, which can be captured using concept mapping activities. Concept mapping incorporates higher-order cognitive skills (Briggs et al., 2016; Canas et al., 2017) because students assimilate new knowledge into their existing framework of knowledge, and not just memorizing facts, which is considered to be a low-level cognitive skill (Momsen et al., 2010). Concept mapping therefore has been successfully used in classroom learning for mapping knowledge (Hay et al., 2008; Pearsall et al., 1997); assessing student understanding in molecular biology, microbiology and other biological science courses (Briggs et al., 2016);

in classroom and laboratory (Kaiser, 2010), and in novice-expert studies (Briggs et al., 2016; Kinchin, 2001a; Chi and Roscoe, 2002; Lee et al., 2022; Mintzes and Quinn, 2007; Price et al., 2021). Due to its significance, the reliability and validity of concept mapping as an effective tool for assessing student learning have been already established in science education literature (Ruiz-Primo et al., 2001).

The biennial concept mapping conferences focus on themes of assessment, teaching-learning, innovative methods, collaborative learning, etc. However, a recent literature review of concept mapping studies indicates the exploration of the tool for open-ended approach of analysis (de Ries et al., 2022); in deep learning techniques for comparing concept maps by automated computation based on similarity (Montanaro et al., 2022); in automatic concept map generation from natural language to derive semantically rich graphs (Lu et al., 2023). This indicates the trend of focus on semantics in knowledge graphs.

Our work too focuses on semantics but in representing scientific knowledge in textbooks with special emphasis on linking words or relations (predicates). This is of significance as the linking words hold explanatory power for describing statements that take part in part-whole, class-subclass, spatial, function, cause-effect, or processes dimensions. Recent studies have also provided evidence about emphasizing linking words (linking phrases) in concept maps. A research study compiled a list of linking words based on their reflections on teaching and extracting the linking words from books to include in undergraduate introductory biology courses describing quantitative, structural, and temporal statements (Javonillo and Martin-Dunlop, 2019). In another study, pathophysiology students created concept maps on respiratory systems and endocrine systems using a standardized set of linking words grouped as dynamic, static, illustrative, definition, and information (Fonseca et al., 2021). The quantifying of linking words resulted into higher number of dynamic linking words, indicating the prominence of processes.

Another prominent area of interest is about studies in novice-expert comparison using concept mapping. Research on expertise studies indicates that experts use a variety of linking words for appropriate meanings. In contrast, novice's usage of linking words seem often inappropriate, and use of similar linking words for a variety of meaning results in ambiguity (Kinchin et al., 2000). As one of the characteristics of expertise is focusing on links, we adopt concept mapping with specific set of linking words facilitating students in expert-like thinking for developing science vocabulary. Studies indicated that providing linking words along with concepts during concept mapping helps to depict more valid propositions as compared to just providing concepts (Kharatmal and Nagarjuna, 2009), helps in facilitating precision and rigor, attaining expertise (Kharatmal and Nagarjuna, 2008). The specific set of linking words with concepts serve as scaffolding in classroom learning (O'Donnell et al., 2002).

Although the concept mapping as a research tool was mostly studied in biology education, it has also shown to be studied widely in other physical sciences such as chemistry education and physics education. Studies have been conducted on using concept mapping in chemistry education on topics such as: atomic properties, states of matter, acid-base, electro-chemistry, etc. A recent study showed learning achievements of using concept maps with fill-in-nodes on students of a general Chemistry course who had low prior knowledge (Huynh and Yang, 2024).

Further, concept mapping with three different modes: static, fill-in-concepts, fill-in-labels, were applied to investigate the effect of learning chemistry in undergraduate classroom. Their results indicated that the fill-in-nodes, fill-in-labels format were found to be influencing learning, where in prior knowledge played a significant role in enhancing learning performance.(Wong et al., 2021)

Studies have also been conducted on using concept mapping in physics learning, physics teacher education, on topics of light, particle physics, etc. A study to investigate effect of concept mapping on physics achievement in comparison with traditional lecture based teaching, was conducted on the topic of electricity. The results showed that concept mapping was more effective than traditional instruction in improving physics achievement (Karakuyu, 2010).

A study showed positive effect of using concept mapping in comparison to the traditional lecture based tutorial strategy in teaching circular and rotational motion (Luchembe et al., 2014). In the domain of physics teacher education, a study determined the effect of reading of concept mapping on the topic of light to convert into meaningful text. This text was then compared with experts' text. The researchers found that over 75% of teacher candidates were successful in converting the concept map into meaningful text (Usta et al., 2014).

2.3 Representational Redescription Model

The Representational Redescription (RR) model is about the gradual transformation of implicit to explicit knowledge in the course of cognitive development, learning, knowledge acquisition as put forward by Karmiloff-Smith (1995).

Knowledge acquisition into the mind can occur in three epistemological ways: (i) knowledge is already innately specified; (ii) knowledge is acquired through physical and socio-cultural/linguistic environments; (iii) by "an internal process of representational change whereby the mind enriches itself from within by re-representing the knowledge that it has already represented" (Clark and Karmiloff-Smith, 1993, p. 495). This epistemological source of knowledge is of interest to Karmiloff-Smith (1995, 1991), that is by the way of *Representational Redescription* (RR) model for cognitive development. In this model, the knowledge acquisition, either innate or acquired, is stored by re-representing recursively its own internal representations. The RR model suggests four levels – Implicit, Explicit 1 (E1), Explicit 2 (E2) and Explicit 3 (E3) – during the process of representational change (Karmiloff-Smith, 1991, 1995).

1. Implicit: The first level-I is implicit, in which knowledge is activated as a response to external stimuli. At this level, the knowledge maps, or procedural knowledge, are unavailable as a data structure. The level-I representations remain bracketed or fragmented without forming interlinks to other domains. Therefore, the level-I

representations remain unavailable to other parts of the cognitive system.

2. Explicit E1: The level-E1 as an explicitly defined representation that can be related to other representations. Level-E1 are redescriptions of abstractions in a higher level language, and at this level, these are open for links with intra-domain and inter-domain. Even though level-E1 representations are available as data to the system, these are not yet necessarily available for conscious access and verbal reports.

3. Explicit E2: At this level-E2, the data are available for conscious access.

4. Explicit E3: The level-E3 involves reconciliation of internal representations and external data that are now available for conscious access and verbalization.

In this way, the RR model, through I-E1,E2,E3, involves gradually redescribing a procedurally based knowledge into more explicit, consciously accessible, and verbal formats. The RR model argues for domain specificity, although the process of Representational Redescription within each domain would be the same. According to Karmiloff-Smith (1991), a domain is a set of representations of a specific area of knowledge, such as biology, language, physics, art, music, etc. Within the domains, are subsets called microdomains, such as cell biology, noun acquisition, gravitational force, etc. (Machado and Braga, 2019).

The RR model was developed to critique the theory of modules (Fodor, 1983) and age-related stage-theory model of cognitive development (Piaget and Duckworth, 1970). In the area of representational theory of mind, the work by Fodor (1983) (1983), suggests the mind would be formed of “modules” or innate input systems, which are encapsulated and functioning independently that are unavailable to the other parts of cognitive systems. However, Karmiloff-Smith (1991) not only rejects that mind is innate with pre-specified modules but instead proposes a gradual developmental of modularization of mind. The RR model argues for a gradual process by which the implicit knowledge becomes explicit and is available for conscious and verbal access to the same knowledge. The representations embedded in the brain that are implicit are redescribed thereby becoming knowledge to the brain allowing for these redescriptions to be available for cross-domain (Karmiloff-Smith, 1991). In this way, representations of an expert’s knowledge emerge over a period as a function of *repetitive refinements* (Mack and Robinson, 1991, p. 265).

The validity of the RR model as a general model of cognitive development has been studied and evaluated for its utility, generalizability, and empirical support (Butler, 2008; Pine and Messer, 2003). The RR model has been considered as a reference for conceptualizations of models to enable self-enrichment in science education (Machado and Braga, 2019). Studies exploring children's representations from implicit to explicit levels in domains of balance and numeracy reported that when children acquire explicit verbal knowledge, they can access it for other tasks within a domain, thus indicating the generalization of knowledge across domains. This generalization process as procedural information becomes redescribed into a verbal and accessible format (Butler, 2008). Researchers providing empirical support to the RR model have suggested modifications by introducing transitional representational levels capturing gradual change from I-E1, E2, and E3 levels. As part of a study on balance tasks with children, these transitional levels were introduced, moving between implicit and abstraction nonverbal levels, and abstraction verbal and E3 levels indicating more significant variability in their performance (Pine and Messer, 2003).

In cognitive psychology, the RR model was applied to understand spelling and reading development in literacy and verbal knowledge acquisition among children (Critten et al., 2007). The cognitive mechanisms by which the spelling development occurs from phonological to morphological pattern, via U-shaped developmental curve, with implicit representations becoming more rule-based and further becoming more explicit verbalizable knowledge prior to the E3 level (Critten et al., 2007). The RR model was used as a cognitive framework to device delivery modules as scaffolding in music listening while developing critical skills in listening and sharing their understanding of music from implicit to explicit knowledge (Boechler et al., 2015). Knowledge redescription can support conceptual dependencies between counting, number structure, and calculation triggering through structured activities. The reification or refinement of the counting process can occur due to their knowledge of oral counting sequences getting redescribed in explicit and accessible formats, as the RR model postulates (Voutsina, 2012).

The RR model, however, is not without any criticisms about its mechanisms of transformations of redescriptions and bears challenges underlying representations at the implementational level, etc. Researchers are critical of how the model is used as an architecture for redescribing in the system (Carassa and Tirassa, 1994). In the natural language domain, the distinction between the E2 and E3 explicit levels seems blurry because the RR model is about how it is employed in the system but not how it is represented. The latter has been part of our research work. Procedural learning can occur when the learner explores a new domain, but it does not imply a representational system. To critically point out that, even though the RR model prescribes the three levels of redescriptions, it does not attempt to explain the mechanisms of these transformations of redescriptions. In summary, the modules, delivery courses, structured activities, constraints, processing, and introducing transitional levels can serve as scaffolding for these mechanisms. We ask: what if our proposed method of representation applying constraints can serve as one such explanation of the mechanism, or can serve as scaffolding between implicit and explicit levels/knowledge?

We agree with Hollan et al. (2000) and Karmiloff-Smith (1991) position about implicit to explicit knowledge being a gradual process, and we strive towards a common or a middle ground by blending these two positions. We, too, consider the implicit and explicit not as a dichotomy but as interfacing in the continuum with constraints and re-representations. Our interest in this area is designing a representation method that can facilitate an interface for transforming implicit knowledge towards explicit knowledge.

In this thesis, Karmiloff-Smith (1991) model of phases/levels of Representational Redescription (RR) is used as a theoretical framework to design and develop a method for re-representation following *repetitive refinements*. We are adopting the RR model as one of the theoretical frameworks because it is about representational changes through which implicit knowledge becomes explicit knowledge. We developed the Re-representation Method as phases based on the previous one in recursive iteration. The theoretical framework was found to be suitable for our research because the thesis aims to focus on making links (linking words or linking phrases) explicit that are otherwise implicit but intended. We echo with Karmiloff-Smith's submission of each Phase as not just a "simple duplication of lower levels, but rather, it involves increasing explicitation and accessibility" Karmiloff-Smith (1995). The common concern in most of these approaches, by linguists, philosophers, logicians, or ontologists, is to bring the implicit semantics to the foreground by making these explicit.

2.4 Nominalization

Systemic Functional Linguistics (SFL) is the study of the relationship between language and its functions in social settings (Halliday, 1985). The SFL follows a stratification model of a linguistic system comprising of meaning (semantics), sound (phonology), and lexicogrammaar (syntax, morphology, and lexis). According to SFL, grammar is treated as a meaning-making resource with an interrelation of form and meaning.

In the context of the language of science, according to the SFL framework (Halliday, 2006; Halliday and Matthiessen, 2004), nominalization of terms is considered one of the benchmarks of scientific text. Nominalization is a process of making (re-representing) relationships (verbs), or qualities (adjectives) into things (i.e. nouns) (Martin and Rose, 2003). Nominalization falls under the grammatical metaphor idea whereby nouns nominalize process verbs without changing the meaning of the text (Halliday and Matthiessen, 2014). Nominalization is also seen as a transition from colloquial to academic or scientific language and has a central role in making the scientific text more concise. Scientific text usually follows this implicitly.

To illustrate how implicit meaning becomes explicit, we quote an example from Halliday (2006, p. 35) and show the significance of nominalization in the language of science.

1. Osmolarity increases, so putrescine is rapidly excreted.
2. Because osmolarity increases, putrescine is rapidly excreted.
3. That osmolarity increases means that putrescine is rapidly excreted.
4. Osmolarity increasing leads to putrescine being rapidly excreted.

5. Increasing of osmolarity causes rapid excreting of putrescine.

6. Increase of osmolarity causes rapid excretion of putrescine.

7. Increases of osmolarity cause rapid excretions of putrescine.

All the above sentences express the same meaning and truth value. However, the language of science evolved to describe and ascribe causal relations between the two phenomena, which became explicit in the latter sentences. The verbs and adjectives used in the former sentences have been nominalized as class terms in the latter sentences. Without getting into the grammatical characterization of the transformations, which are detailed by (Halliday, 2006, p. 84-85), let's focus on how the expression "osmolarity increases" becomes "increase of osmolarity"; "so" becomes finally a "cause" via "because", "effect", "leads to"; how the adjective "rapidly" becomes the noun "rapid"; and the verb "excreted" becomes the noun "excretion". After such repeated transformations, some of the verbs and adjectives eventually become nominalized class terms, leaving the semantic relation we extracted to express causation. Using prepositions to relate class terms to create condensed expressions, as in "increase of osmolarity" and "excretion of putrescine" must not be ignored. Such transformed usage of verbs and adjectives into nouns makes the language of science difficult to learn and is widely documented in science education and linguistics literature. Our interest in this context is to find what verbs remain after these re-representations. We notice that those semantic relations are precisely part of the various and growing published ontologies and the fundamental logical relations that remain. The remaining and *distilled* semantic and logical relations are the common ground of everyday language, scientific language, and ontologies.

In scientific language, these semantic relations need to be well-defined for disambiguation. This is furthered by adopting well-defined relations (object properties and data properties) from the published biomedical ontologies.

2.5 Biomedical Ontologies

Biomedical ontology research and development work is a proliferating field with interdisciplinary researchers from computer science, philosophy, data and software engineering, logic, linguistics, medical practitioners, etc. (OBOFoundry, 2020). The need for the research area of creating biomedical ontologies arose from the post-genomic era due to the shift from mapping genomes to analyzing information by applying controlled vocabulary. Biomedical ontology research involves entities' specifications, attributes, and relations among them (Rubin et al., 2008). These are controlled vocabularies describing the semantics data in a human and machine-readable way, widely used for processing information in biomedical research and healthcare systems (Konopka, 2015). Biomedical ontology, in contrast to biomedical terminologies, goes beyond mere dictionaries of biological terminologies or biological products to controlled vocabularies, to principled knowledge structures, to applied reasoning about biological knowledge (Bodenreider et al., 2005). Creating a database of biomedical terminology comprised of names of substances, qualities, and processes within the biomedical domain. On the contrary, the purpose of biomedical ontology is to develop a principled definition of biological classes and the relations among them. One such instance of this research is the Open Biomedical Ontology (OBO) Foundry, a community-based initiative for developing new ontologies and reforming existing ones (OBOFoundry, 2020; Ontology, 2020). Recent works on library of generic concepts have been useful resource (Barker et al., 2001).

Various initiatives exist in this area of research and are growing, especially in the area of biomedical and healthcare research. The Unified Medical Language System (UMLS) aims to merge existing vocabularies. The Foundational Model of Anatomy (FMA) represents the classes and relationships for modeling the structure of the human body (Rubin et al., 2008; Rosse and Mejino, 2003). Such curated ontologies are available through the public domain, for example, Ontobee portal (Ong et al., 2017).

One of the essential concerns in the biomedical ontology research and development work is the disambiguation of classes and relations (Smith et al., 2006). This is parallel to the concerns we raise but about the language used in the text. Within ontology research, biomedical ontology is concerned with the principled definition of classes and relations between them (Bodenreider et al., 2005), to provide consistent and unambiguous formal definitions of the relations in the ontology, to enable automated reasoning about spatio-temporal dimensions in biological and medical phenomena (Smith et al., 2005). The domain of biomedical ontology modeling introduced the ontological framework for systems biology (Podkolodnyy and Podkolodnaya, 2016) and experimental sciences (Hafner and Fridman, 1996; Fridman Noy and Hafner, 2000) that is relevant for this thesis for representations of text in science education. Our thesis, therefore strives towards a novel work because we focus on the nature of semantic relations in statements in science text, which are usually taken for granted as these belong to everyday language vocabulary.

One may ask, when the biomedical ontologies are already curating semantically well-defined relations, then what is the uniqueness that we bring into our research program? We argue that the biomedical ontology community focuses on curating and defining the class terms and relations. Therefore, we borrow the well-defined relations for re-representing textbook knowledge and thereby creating the Reference Set.

2.6 Conceptual Spaces

The Conceptual Spaces framework for Knowledge Representation has been immensely influential within the fields of cognitive science, cognitive linguistics, artificial intelligence, semantic web, etc. (Bechberger and Kühnberger, 2017). A conceptual spaces framework comprises many quality dimensions that are based on perception, built up from geometrical representations (Gardenfors, 2000; Gärdenfors, 2004, 2011). The "quality dimension" (also known as variables in model-based reasoning) has explicitly been significant to our work on re-representing attribute space as part of developing the re-representation method.

According to Gärdenfors (2004), quality dimensions, such as temperature, weight, brightness, pitch and the spatial dimensions are of height, width and depth as they are produced by sensory receptors (Gardenfors and Warglien, 2012). The brightness dimension is perceived by the visual sensory system, pitch by the auditory system, temperature by thermal sensors, and weight by kinesthetic sensors, respectively (Gardenfors and Warglien, 2012). In addition, there also exist other quality dimensions that are non-sensory. In the Conceptual Spaces framework, the primary function of the quality dimensions is to represent “qualities” or “properties” of objects and the relations between them (Gärdenfors, 2004, p. 11). One may ask how quality dimensions are related to representations in scientific language? At this point, Gärdenfors (2004) makes a clear distinction of quality dimensions as phenomenal and scientific interpretations. The distinction is between those based on perceptions, memories, etc., and those part of scientific theory. The phenomenal interpretations of the quality dimensions provide semantics for a natural language, which is also of interest for our work.

Ontologically, some quality dimensions are considered innate, for example, the sensory dimensions (Gärdenfors, 2004; Quine, 1948). As the learning process occurs during cognitive development, more dimensions are gradually added. For example, in the conservation task (Piaget and Duckworth, 1970), children initially think only about the “height” of the liquid in the container. Later, these develop to separate the dimension of “volume” from the “height” of the liquid, which is the relationship of the other two qualities. Additionally, some different dimensions may also be culturally identified. For example, the time dimension is circular in some cultures. A socio-cultural interaction wherein people share knowledge may also evolve into conceptual spaces.

Further, Gärdenfors (2004), emphasizes the quality dimensions in science—for example, the distinction between “weight” and “mass” in Newtonian mechanics. Although the senses in human perception may not make this distinction, these are learned while adopting the conceptual space of Newtonian mechanics. These distinctions either between “volume” and “height” or between “weight” and “mass” can be understood as “shifts of conceptual spaces” much like the “paradigm shifts” in cognitive development, and scientific revolution (Kuhn, 1970).

To summarize, we develop a Re-representation Method, with the insights from the

'Knowledge as a Network' model, that bears significance in the semantic relations or links in a statement. We focus on the nature and kinds of the links. This can help in making the implicit meaning explicit as we re-represent the statements by following the 'Representational Redescription' model. How do we do it? We do it by borrowing the well-defined semantic relations from the biomedical ontologies, process nominalization from the SFL, predicates from conceptual spaces, to create a Reference Set.

At this point, we bring connections to the research program for creating a "nuclear vocabulary" (Stubbs, 1986). Complementing to this, we explore a research program for creating a "nuclear predicate vocabulary" or "Reference Set" for its use in the language of science.

2.7 Identifying Research Gaps

Graphical representation tools such as concept maps and mind maps are widely used in education. However, their focus was mainly on organizing the concepts (vocabulary) in a hierarchical structure with little or no focus on using well-defined relations, leaving scope for ambiguity. Problems associated with lack of rigor in concept mapping are reported (Sowa, 2006; Kharatmal and Nagarjuna, 2006b; Kremer, 1995). In contrast to concept mapping, SemNet attends to semantic relations more explicitly (Fisher, 1990, 2002; Fisher and Kibby, 1996; Fisher et al., 2000).

We identified certain inconsistencies while assessing the concept maps from published sources. We state that the linking words that provide meaning to the concepts are ambiguously and inconsistently used. Here we borrowed the insights from the classical work on nature, usage, and definitions of semantic relationships (Winston et al., 1987) and applied for disambiguation in maps. Some of these are *meronymic inclusion, class-subclass inclusion, spatial inclusion, attributes,* etc.

2.7.1 Problems in Accuracy of Linking Words

We began with studying the published concept maps in the public domain, journal articles, and books. The findings of concept mapping showed that it can be used for meaningful learning, conceptual change, etc.(Novak, 1990; Martin et al., 2000). However, with a deeper analysis of these works, we identified certain inconsistencies in the nature and usage of the linking words in the standard concept mapping and scoring methodology. The inconsistencies may not be a problem in other subjects like physics, chemistry, and mathematics, as this problem does not arise due to the use of formulae that use a minimal set of predicates. Since the language of biology cannot be formulated in such a language, we suggest that a controlled natural language can be used instead. Our research work can demonstrate its possibility. We proposed certain refinements of linking words in terms of accuracy and validating hierarchies in concept mapping.

The concept map on a focus topic of "life in the ocean" (Martin et al., 2000) is shown in Figure 2.1. In this concept map, S1 depicts using the linking words: 'consists of', 'have 2 groups', 'are either' and 'include'. A deeper analysis of the S1 suggests that only the first level is part-whole relation while all the other levels follow a class-subclass relation (Winston et al., 1987). We refined these linking words and re-represented those with the semantically well-defined linking words as shown in S2. By this, we suggest weeding out ambiguity, bringing in consistency i.e. to use the same kind of linking words for the same meaning throughout the concept map.

Another gap we found was about invalid criteria for counting hierarchy levels. We argue that the criteria to assign scores for the hierarchy are graphical-based and not semantic-based. Logically, in a well-formed hierarchy for representation, the relations are transitive, asymmetric, using a single linking word consistently (Mayr, 1969). In S1's concept map the first relation takes part in the part-whole relation and all the other relations in the hierarchy take part in the class-subclass relation, though S1 uses different linking words. Therefore, the standard scoring method does not consider this subtle distinction and therefore depicts a logical fallacy. This is a serious anomaly. Given this context, there was a need to find an appropriate refinement for validating hierarchies in the concept maps which is the use of same linking word for same meaning through the same hierarchy.

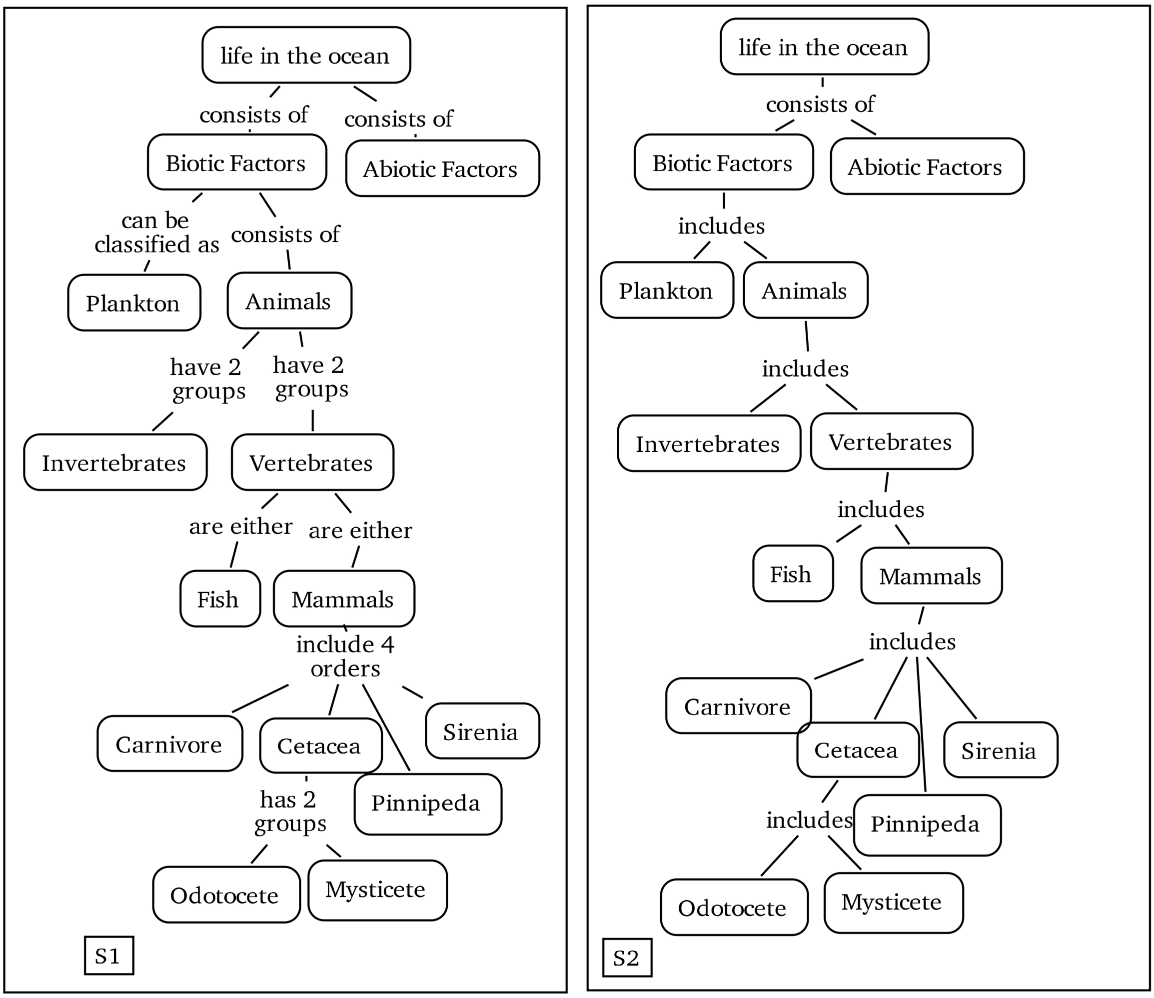


Figure 2.1: Concept map depicting ambiguities and inconsistencies in linking words and suggesting refinements.

Through the re-representation, we could not only refine the linking words for accuracy but also help in validating hierarchy in concept maps (Kharatmal and Nagarjuna, 2006b).

2.7.2 No distinction between relations and attributes

To indicate the lack of distinction between relations and attributes, we illustrate from published literature and suggest its re-represented version. Consider the propositions:

(1) [sharks]→(can be)→ [great white shark, tiger shark]

(2) [shark teeth]→(can be)→ [big, small]

These propositions use the linking words "can be" ambiguously to mean two different sense. It is used to represent a class-inclusion relation, and also for attribution relation. The corresponding re-represented propositions in the same order are:

(1′) [sharks] → (includes) →[great white shark, tiger shark]

(2′) [shark teeth]→ (size can be)→ [big, small]

We eliminated ambiguous relation name *can be* by resolving it into *includes* in (1′), *has size* in (2′). As part of re-representation, we assign variable properties as attribution. This distinction becomes necessary when describing processes and objects in terms of quantifiable predicates, which is required for model-based reasoning. We assume in this work that model-based reasoning and rigor are epistemic requirements for practicing science, and therefore it is normative for introducing this explicitly in science education.

2.7.3 Lack of focus on process-centric propositions

Using verbatim linking words (verb based) in textbooks such as 'produces,' 'supports,' 'stores,' 'secretes,' etc. gives a colloquial sense to scientific propositions (statements). It eliminates the focus on processes in biology. This requires re-representation to bring the focus to the process-centric proposition. The concept mapping community also shares the concern, that the focus question is not just "what" but "how" to facilitate dynamic propositions or process-centric propositions (Miller and Canas, 2008). In order for the processes to come into focus, we need to nominalize (Halliday and Matthiessen, 2004) these action-verbs and re-represent them as processes as depicted in the re-represented propositions.

For example, in the sentence, "endomembrane system regulates protein traffic", the predicate "regulates" when re-represented, is nominalized into the process as "regulation of protein traffic". The nominalized process terms bring the focus of the objects to place "protein traffic" into processes such as "regulation of protein traffic" and create dynamic propositions such that the representation is to understand the "regulation of protein traffic" as a biological process vis-a-vis mere "structure of protein". This can enable to focus on the question of 'how' in understanding any phenomenon or biological process.

2.7.4 No focus on depicting change

Although concept mapping has been used almost worldwide, researchers have pointed out the need to deploy at its full potential. The concern is that the tool is used primarily for depicting descriptions and not yet for representing explanations, change, or dynamic systems (Canas and Novak, 2006). One of the reasons considered is focus questions that are more description-based than explanation-based. It is also reported that the concept maps have a strong advantage of representing static relations between concepts. Still, the methodology lacks the potential for representing change or dynamic relations (Derbentseva et al., 2007). This indeed calls for re-thinking the method to extend the concept mapping methodology to its full potential, mainly because science is essentially about model-based reasoning and modeling change. In this context, we suggest to use the variable properties (attributes) for modeling change.

Through the above illustrations, we identified the problems and gaps in mapping a knowledge domain. Ambiguity in linking words or linking phrases and lack of rigor in concept mapping is already been reported by other researchers (Costa et al., 2004). In the Artificial Intelligence (AI) community, concept maps are considered less rigorous in methodology (Kremer, 1995; Sowa, 2006), lacking in Knowledge Representation (KR) formalisms (Canas and Carvalho, 2004).

Using linking words for precision and rigor in scientific language accurately characterizes expertise (Kharatmal and Nagarjuna, 2008). Studies indicated that providing linking words helps to depict more valid propositions than just providing seed concepts with the freedom to use any linking words (Kharatmal and Nagarjuna, 2009). In the early attempts, the focus on linking words was studied by Fisher et al. (2000) to create semantic networks as an extension of concept mapping activities using computer-based software. It introduced the software for knowledge organization only. In the novice-expert studies, experts use a variety of linking words appropriate for specific meanings, while novices' usage of linking words seem often inappropriate, and using similar linking words for a variety of meanings, thereby resulting in ambiguity (Kinchin, 2001b).

As the goal of mapping knowledge is to facilitate meaningful learning, it is essential that the meaning providing linking words needs to be unambiguous, consistent, and parsimonious. This can be achieved by replacing the vague and ambiguous linking words with semantically well-defined predicates. This we call *Re-representation*. The predicates are re-represented from the validated and open biomedical community of experts repository (OBOFoundry, 2020).

Our work tries to bridge these gaps and hence our focus is on re-representing the linking words to make it explicit and rigorous and studying if it can bear significance in achieving expertise and restructuring of knowledge. To address the gaps, we framed four research questions (RQ).

2.8 Research Questions

- RQ1: Can we design a method for representation of biology language focusing on the semantics of predicates? What features of biology language could emerge from it?
- RQ2: Could this method be used for the semantic content analysis of biology language? What would the analysis unravel about the nature of biology knowledge?
- RQ3: Could this method be feasible? To what extent using the finite set would be effective in representing biology knowledge?
- RQ4: In what ways would the semantic content analysis serve as a measure of expertise? Could this method be used as proximity indicator?

To address RQ1, we begin with designing and developing a method by re-representing linking words that lead to the modeling of biology language (chapter 3). To address RQ2, we use this method to demonstrate its use for the semantic content analysis of biology textbooks and students' descriptions with procedures (chapters 4 and 5). RQ3 is addressed by demonstrating the method's feasibility and efficacy by comparing it with traditional methods (chapter 6). RQ4 is addressed by comparing the text with the expert's knowledge base and expert's ontology knowledge base to check if the re-representation of linking words can show a proximity percentage (chapter 6). Chapter 7 discusses the conclusions and implications of the work, followed by contributions and the future scope of the thesis.

2.9 Summary

This chapter discussed the theoretical frameworks for developing the method. In Chapter 3, we establish connections between the theoretical frameworks for developing the method.

Chapter 3

Design and Development of the Re-Representation Method

3.1 Introduction

In this chapter, we discuss the design and development of the Re-representation method for semantic content analysis of scientific text. As mentioned, our early explorations have led us to develop the re-representation method at each phase through a continuous re-iterative reflective inquiry process to test and refine the method parallel to a Design Based Research (DBR) paradigm.

The DBR is an amalgamation of design, research, and practice co-occurring in a synergistic relationship with designing and researching (Wang and Hannafin, 2005; Wang, 2020). According to Wang and Hannafin (2005), DBR is "a systematic but flexible methodology to improve educational practices through iterative analysis, design, development, and implementation, based on collaboration among researchers and practitioners in real-world settings and leading to contextually-sensitive design principles and theories". One of the critical characteristics of design experiments has been about "conducting a rigorous and reflective inquiry to test and refine innovative learning environments as well as to define new design principles" (Brown, 1992; Collins et al., 2004; McKenney and Reeves, 2018).

We appropriate the DBR into our designing and development of the Re-representation Method.

Initially, we focused only on a preliminary set of 7 types of linking words belonging to *meronymic inclusion, class inclusion*, and other relations. We tested this by deploying it for textbook analysis. This Reference Set catered for representing structure terms in the topic. As we analyzed more textbooks from higher levels of complexity, the Set was necessary but insufficient. We observed a need for more linking words, for example, *spatial inclusion*, and *function*. This called for iteration of the re-representation method to include other types of relations (iteration 1). In addition to the relations, we represent attributes or variable properties distinctly (iteration 2).

Further, as we analyzed higher-class textbooks, we needed to analyze dynamic propositions (iteration 3) for process-centric statements. Once the process terms were brought to focus, by applying the variable properties, we re-represented change for depicting process modeling for scientific explanation (iteration 4). The method further re-represents the experimental procedure (iteration 5) and linking propositions for inferences (iteration 6). The re-representation method has evolved through features unique in design practices such as creating, testing, iterative refinements, and continuous evolution of the design (Anderson and Shattuck, 2012).

3.2 Research Model

The graphical representation summarizes the overview of the research model complementing with theoretical frameworks and approaches and is presented in Figure 3.1.

3.3 Thesis Objective - 1

We elaborate on this development of the re-representation method with testing, refining it as an iterative process in six phases, and illustrating each phase with examples. We emphasize that the Phases 1-4 have been part of the semantic content analysis, and

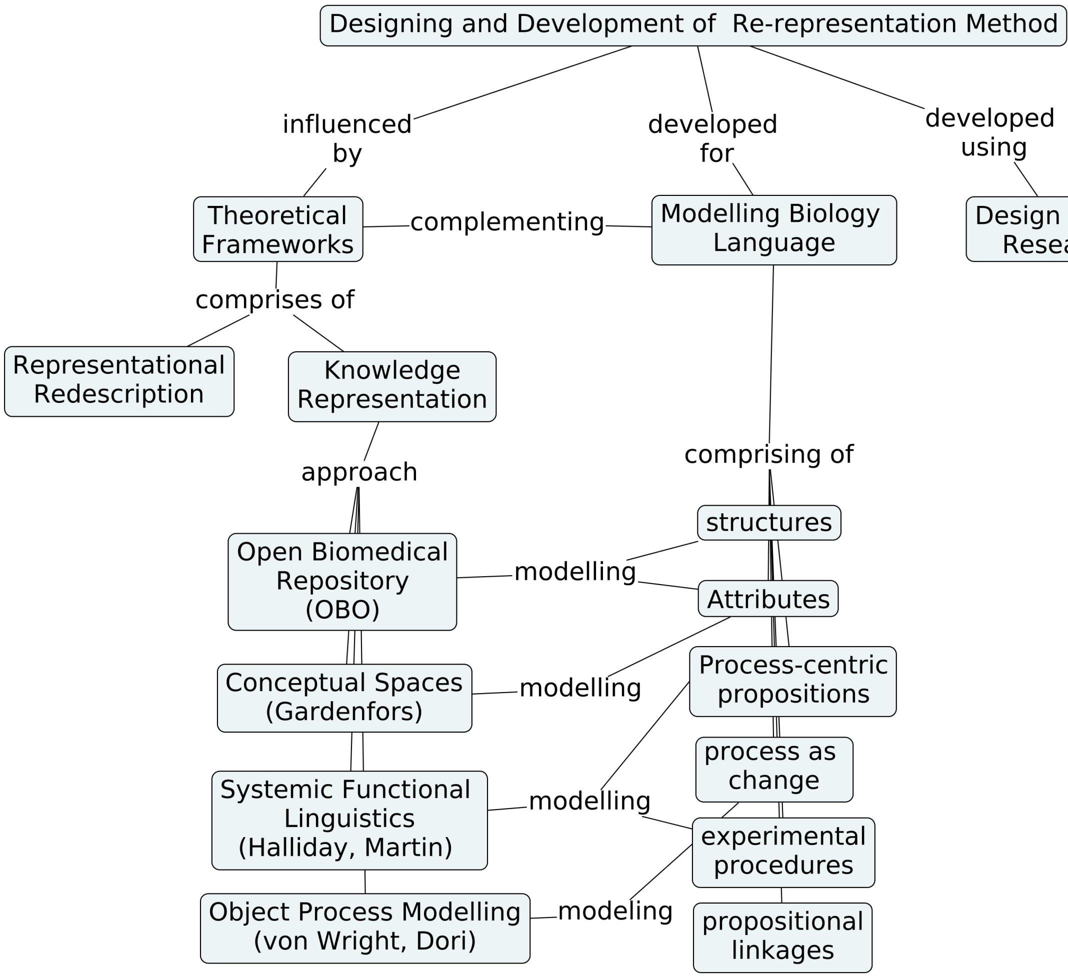


Figure 3.1: A graphical representation of the research model followed for research question 1. (To read from top to bottom and left to right).

evaluation studies. The Phases 5 and 6 serve as an extension of the method for illustrating representation of experimentation and not for any evaluation purposes.

- Re-representation Phase - 1: Re-representing the relations (linking words) (applying open biomedical ontologies, library of generic concepts)

- Re-representation Phase - 2: Distinction of relations with attributes (applying open biomedical ontologies, library of generic concepts, conceptual spaces)

- Re-representation Phase - 3: Re-representing dynamic propositions (applying nominalization, process specification language, library of generic concepts)

- Re-representation Phase - 4: Re-representing process as change for process modeling (applying Knowledge Representation, Object Process Modeling, process specification language, library of generic concepts, open biomedical ontologies, conceptual spaces, nominalization)

- Re-representation Phase 5: Representing experiments (applying the SFL framework)
- Re-representation Phase 6: Propositional linkages for inference-based statements

A note: These are not exhaustive phases. It is possible that, if and when the method is applied to another domain, the phases may gradually increase or evolve.

3.4 Re-representing Using a Constrained Set - Zeroeth Version

As we discussed in the early explorations, we commenced the journey of focusing on the nature and kinds of linking words that were used in concept maps, by referring to the classic work on the taxonomy of semantic relations (Winston et al., 1987). Mapping a chapter on cell structure and function from class 8 textbook (NCERT, 2007b), we arrived at 7 linking words belonging mostly to meronymic and class inclusion relations. However, as we further analyzed class 9 (NCERT, 2007c), there was a need for other dimensions such as spatial inclusion and function. The set was added with more types of linking words while analyzing higher-level textbooks.

Although concept mapping methodology had been our initial reference for concepts and linking words, we suggested the replacement of ambiguous linking words with well-defined linking words. As the work was on textbook analysis, we focused on statements. According to concept mapping, two concepts connected with linking words form a proposition or a statement. The concept mapping and textbook have a standard format for representing propositions (statements). We shall refer to the linking words as relations because we wanted to represent a standard format (not just for concept mapping alone). Henceforth, we shall refer to the linking words or linking phrases as relations (or use them interchangeably in the context of concept mapping work).

3.5 Re-representing the Linking Words – Phase-1

The principle objective for the re-representation method has been to disambiguate the linking words, either used in concept mapping or in textbook sentences and re-write those with more explicit and well-defined relation terms. Let us begin with an illustration. Consider the sentence, "Cell has nucleus" (A). The intended meaning of the sentence A could mean any of the following: "Cell contains nucleus" (B); "Cell consists of nucleus" (C); "Cell constitutes nucleus" (D); "Cell composed of nucleus" (E); "Cell is made of nucleus" (F); "Nucleus is inside the cell" (G); "Cell possesses nucleus" (H).

How do we determine the intended meaning of sentence A? Without additional specifications, it is not possible. The later sentences B to H provide alternatives, but all of them may or may not be correct. What one wants to express is based on the choice of the linking words used between the subject and the object of the sentence in the given context. In the above sentences, the subject terms and object terms are constant, though the position of the subject and object may be changed as in sentence G. From this example, we classify the later sentences as having more clear expressions than sentence A, which used a vague and ambiguous linking word "has". This has been re-represented with "consists of" as the intended meaning of the sentence is meronymy inclusion relation. Therefore the choice of the linking word/ relations depends on the intended meaning that requires the linking words to be re-written to make the meaning explicit and unambiguous. As "has" can be used to express several meanings, but "consists of" has to be used to express only one meaning i.e. meronymy.

We now take a few examples from the textbook to illustrate the re-representation (Phase-1) of the linking words and disambiguate it with semantically well-defined relations. This requires a few steps:

- Step 1: Extracting and paraphrasing the sentences into propositions
- Step 2: Re-representing the propositions

Each of the above steps is an instance of the Representational Redescriptional model

(Karmiloff-Smith, 1991, 1995) and therefore helps in making meaning explicit at each step.

Content Sample - 1: Mitochondria

We discuss a passage from Class 9 text on Mitochondria (NCERT, 2007c, p. 57). Below is the verbatim passage:

> Mitochondria are known as the powerhouses of the cell. The energy required for various chemical activities needed for life is released by mitochondria in the form of ATP (Adenosine triphopshate) molecules. ATP is known as the energy currency of the cell. The body uses energy stored in ATP for making new chemical compounds and for mechanical work. Mitochondria have two membrane coverings instead of just one. The outer membrane is very porous while the inner membrane is deeply folded. These folds create a large surface area for ATP-generating chemical reactions. Mitochondria are strange organelles in the sense that they have their own DNA and ribosomes. Therefore, mitochondria are able to make some of their own proteins.

3.5.1 Paraphrasing the Sentences into Propositions

. We extract each sentence from the passage and show its paraphrased propositions as shown in Table 3.1 and in Figure 3.2. (Please note: The Tables and Figures may appear at the end of each chapter, if not within the text).

The passage with 9 sentences is paraphrased into 20 propositions because the compound sentences are broken into simple sentences to create propositions. A schematic is shown in Figure 3.3.

Usually, a text passage on a topic comprises a number of sentences in the form of simple and complex/compound sentences. The sentences are paraphrased to form propositions. According to Wittgenstein (1922), “a proposition is either true or false”. Usually in natural language, the proposition is in the form of subject–predicate–object.

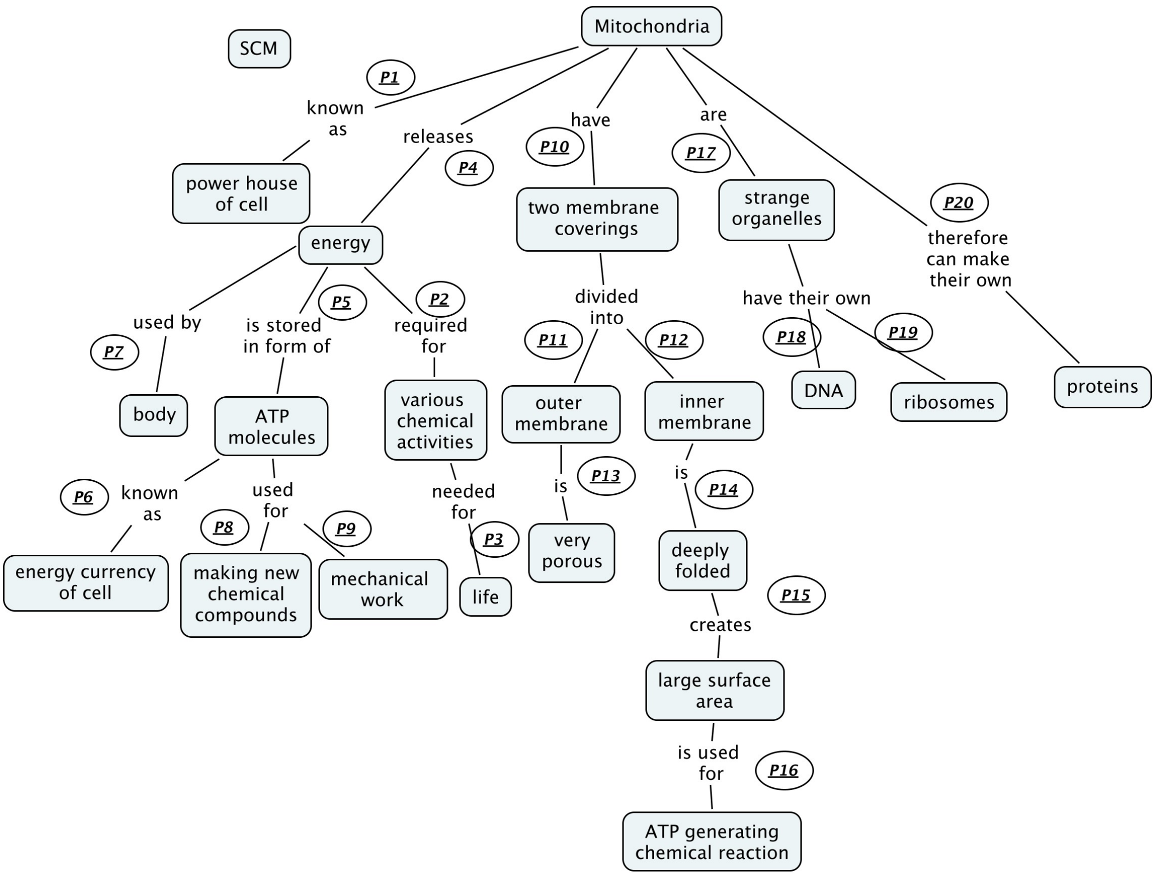


Figure 3.2: A graphical representation of the topic of mitochondria using concept map.

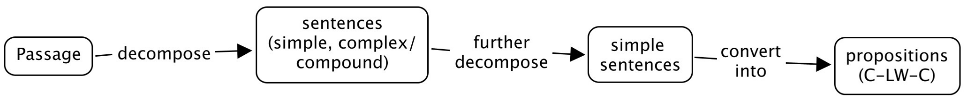


Figure 3.3: Schematic showing the extraction of passage to form sentences and further to form propositions.

In the concept mapping method, its equivalent is concept–linking word–concept form. In knowledge representation (KR), a proposition is represented as a triple in the form of node–link–node.

A simple sentence can yield one or more than one proposition. In this way, paraphrasing from sentences can yield into a number of propositions. The method of paraphrasing is similar to that of extracting sentences from a lexical database into basic propositions, for example, creating Medical Wordnet from WordNet database (Smith and Fellbaum, 2004). We discuss the paraphrasing process for the passage on Mitochondria by referring to Table 3.1.

- Simple sentence yielding single proposition: S1 to P1; S3 to P6; S9 to P20.

- Compound sentence yielding more than one proposition: S2 to P2-P5; S8 to P17-P19.
- Underlining linking words: Linking words of all the propositions yielding after paraphrasing are underlined as shown in Table 3.2.
- Avoiding redundancy: After paraphrasing, P5 is obtained twice from S2 and S4. In both sentences, S2 and S4, the subject of the sentence is "Energy". The predicates of "Energy" are "released in the form of" and "stored in the form of". To avoid this redundancy of meaning in linking words both these sentences are paraphrased and combined using the same proposition, P5. The concept mapping has the advantage of eliminating redundancy or repetitions which are in the text. This convention of concept mapping results in a network representation avoiding the repetition of subject or object terms following the graph /network model.
- Use of inverse relations: Proposition P4 can be paraphrased in two ways by interchanging the subject and object position and choosing the linking words appropriately. Such linking words are called inverse relations. Semantically, the two propositions are equivalent. In P4, the second proposition is selected for the map in order to establish linkage with the pre-existing subject of "Mitochondria". Similarly, the proposition P7 is paraphrased in two ways. The second proposition is chosen because of the prior existence of the subject "Energy" in the map.
- Adding words to make the meaning explicit: In the sentence S4, although the sentence mentions only "for", we paraphrased it as "used for" to make the meaning explicit, yielding P8 and P9.
- Residue: The phrase "instead of just one" used in the sentence S5 has been ignored since concept mapping does not define how to represent qualifiers in the map. This is also as an example to show that a concept map cannot take care of everything that can be expressed in a text form. Let's call this as residue after paraphrasing. Similarly, "in the sense that" in S8 is also ignored.
- Adding a proposition to establish connection: The terms "outer membrane" and

"inner membrane" do not exist in S5, but they do in S6. To represent the connection between S5 and S6, two propositions, P11 and P12, are added between "two membrane coverings" and "outer membrane" and "inner membrane", which will also facilitate the representation of the connections in P13, P14, P15. Thus, concept mapping helps in showing the branching relations between propositions more clearly than in the text form.

- Alternate paraphrasing: In P13 and P14, it is also possible to paraphrase "inner membrane" and "outer membrane" as "porous outer membrane" and "folded inner membrane". Using this alternate paraphrasing, we can now write "folded inner membrane creates larger surface area" instead of P15. Though the subject in P15 refers to a property, therefore the sense of the proposition is not grammatical. This issue will be resolved in iteration 2 of the re-representation method (Phase 2).

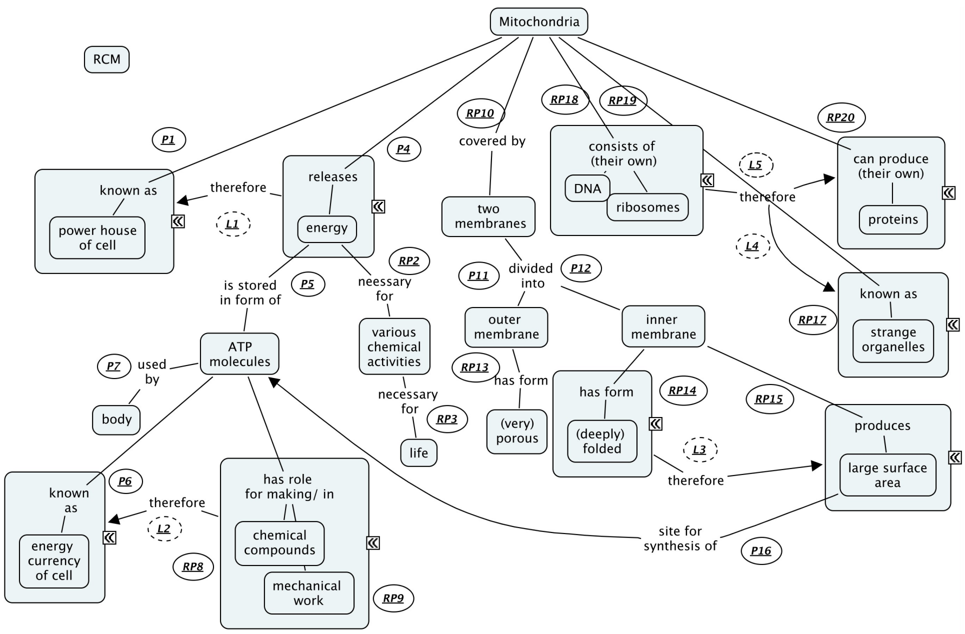


Figure 3.4: Re-representation of the topic of mitochondria using concept map.

We illustrate the re-representation of the same text in Table 3.2.

The re-represented propositions are emphasized in the list. We have ascertained that there is no loss of knowledge, and only when required are the linking words replaced from the Reference Set.

The Propositions P3 and P4 have been re-represented by replacing the linking words "necessary for" as RP2 and RP3, respectively, from the Reference Set. Not only the linking words are replaced but also are used consistently in both propositions. In propositions P8 and P9, we have replaced the linking words with a more specific kind "has role" which denotes that the object (synthesis of new chemical compounds, performing mechanical work) specifically brings focus on the process (more details are discussed in iteration 3).

In proposition P10, the linking word is replaced with a particular kind from "have" to "covered by" as in RP10. The linking word "have" is ambiguous and may denote either a part and whole relation or containment relation (Keet and Artale, 2008; Kharatmal and Nagarjuna, 2010). To avoid this ambiguity, a more precise linking word is used from the Reference Set. As the relation between Mitochondria and the two membranes is of a covering, the linking word "covered by" (Smith et al., 2005) gives accurate sense and meaning to the proposition, in place of "have".

In another set of similar propositions, P13 P14, the linking words are changed from "is" to a precise "has form" as in RP13 and RP14, respectively. We suggest these kinds of linking words to be distinct. We explain this in detail in Phase 2.

In continuation, the linking words in propositions P15 and P16 have been replaced with "produces" and "site for" respectively. Usually, when a map is created from the text, the linking words "creates" and "for" are used. Not that there is any problem. However, a suitable choice of linking words would be "produces" for RP15 to ascertain a cause-effect relation. In RP16, the inner membrane gives a sense of being a site or location for chemical reactions to occur. Hence, a suitable word "site for" is replaced from the Reference Set.

Now, let us consider the propositions RP17 to RP20 (from re-representations). These re-represented propositions are depicted as a re-represented concept map shown in Figure 3.4. While nothing is wrong with these propositions, we went a step ahead. We represented these propositions using the nested node feature of CmapTool to convey the propositional connectives between these three propositions. The bottom part of the re-represented concept map in Figure 3.4 graphically conveys the linking between propositions by using the conjunctions "because" and "therefore". We have re-represented these propositions and created relations between propositions using the same feature of nested nodes in CmapTool.

In the following sections, we explain the justification for re-representing the linking words with semantically defined relations.

3.5.2 Re-representing Propositions

During our course of research work, we noticed that the linking words used were incorrectly used in expressing the intended meaning thus giving rise to ambiguity in the proposition. Although the intended meaning was implicit there was a need to replace these with well-defined linking words so that the meaning becomes explicit. The recent rise of interest in Knowledge Representation (KR) and Artificial Intelligence (AI) research is mainly motivated by projects in computer science and cognitive robotics and also demands semantic relations. This resulted in the creation of libraries of well-defined predicate terms for every domain along with concept terms, generally called ontologies.

An ontology is a formal and explicit description of concepts (classes/objects), their properties (data properties), relationships (object properties) between the concepts. In this work, we would make use of the work resulting from Open Biomedical Ontologies (OBO) (OBOFoundry, 2020), Relation Ontology (RO) (Ontology, 2020), library of generic concepts (Barker et al., 2001). We refer to this library in this work as Reference Set.

As and when we notice inaccurate usage or ambiguous linking words, we replace them with well-defined linking words from the Reference Set. How do we know the linking words are ambiguous? This question cannot be settled without making assumptions about the intended meaning of the text, which cannot arise without domain knowledge and its context. We will illustrate this argument for Re-representation through examples.

Often, linking words such as “is a/are”, “has/have”, “can be/may be”, “in/on” etc., are used in science textbooks. Although position-wise, they are similar to the linking words between nouns or concepts, these provide a colloquial sense when used in scientific language. The problem is these linking words seldom convey the exact or intended meaning and give rise to ambiguity. While these words are not at all dismissed, we suggest their usage with the appropriate linking words, such as using "is found in" instead of using "is in", similarly to using "part of" instead of just using "has". Our approach has been towards resolving this ambiguity as part of Re-representation and replacing them with semantically well-defined relation names (linking words) such as – "part of", "includes", "surrounded by", "located in", "has function", etc. This is done as part of the Re-representation of propositions. We will discuss these situations from the context of the data of content analysis.

The classical paper about semantic relationships (Winston et al., 1987) provided the primary source for Re-representation of linking words. The semantic relationships are organized into *Inclusion (Meronymic inclusion, Class inclusion, Spatial inclusion), Possession, Attachment*, etc. (See Figure 3.5. Each of these linking words is semantically well-defined, and therefore, the inclusion relations mentioned in this classic work have been used during re-representation. Another significant source for a repository of well-defined relations is the state-of-the-art research on Biomedical Ontologies (Smith et al., 2005). During re-representation, we apply the relation names defined in the Relations Ontology (RO) (Ontology, 2020) of the formal knowledge representation group, the Open Biological and Biomedical Ontologies (OBO) (OBOFoundry, 2020). In OBO, the relation names are categorized as *foundational, temporal, spatial, and participation* (Smith et al., 2005). The rationale for choosing relation names from the RO is that every relation name in the Relations Ontology has well-defined semantics, thus making the propositions precise and unambiguous. Below are definitions of some of the critical relations (Smith et al., 2005).

> is a =def. For continuants: C is_a C' if and only if: given any c that instantiates C at a time t, c instantiates C' at t. For processes: P is_a P' if and only if: that given any p that instantiates P, then p instantiates P.'

> 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

contained in =def. C contained_in C' if and only if: given any instance 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' at t, and it is not the case that c *overlaps* c' at t. (c' is a conduit or cavity).

The OBO and RO refer to the relations as object and data properties linked to class terms. As mentioned in Chapter 2, we apply the OBO for curating the relations and thus build the Reference Set. The procedure followed in this iteration is to notice the linking words that are ambiguous and re-write these by replacing them with the terms found in OBO or RO. This would result in a re-represented proposition.

We now explain the rationale for applying the procedure through the following illustrations of the most widely used ambiguous linking words that require re-representation.

"Is-a/Are" Linking Words

The "is-a" (also known as "IS-A") is a primary relation that links class-subclass, forming a hierarchy and enabling property inheritance. The "is-a" relation is used for subset-superset relation, for example, "Car is a vehicle"; "A triangle is a polygon"; "Whale is a mammal". All these propositions, even though they are using the "is-a" relation, have the same intended meaning of class-subclass. Therefore, the use of "is-a" (or "are") is used only to mean class inclusion relations. The KR community explicitly defines the semantic meaning of the "is-a" relation (Soldatova, 2010; Smith et al., 2005) as:

> is a =def. For continuants: C is_a C' if and only if: given any c that instantiates C at a time t, c instantiates C' at t. For processes: P is_a P' if and only if: that given any p that instantiates P, then p instantiates P'.

We can understand the "is-a" relation with the following example: Mammalia (C) is_a vertebrate (C') if and only if, given any instance (c) of Mammalia (C) will also (c) be an instance of vertebrate (C').

The "is-a" relation is commonly used in any science text. Still, it is also used for various kinds of meanings, for example, in meronymy (part-whole), spatial inclusion (location specifying), and instance of (member of a class) relations. The linking word "is-a" giving rise to ambiguity is illustrated with examples in Table 3.3. These popular uses of "is-a" can be replaced with more specific relations in the re-representation that we propose; we reserve "is-a" for class-subclass relation, which follows from the definition given above.

Table 3.3 has verbatim sentences from text passages, their paraphrased propositions, and re-represented propositions. In sentence S1, the usage of the linking word "is-a" giving rise to ambiguity can be noticed. Usually, the intended meaning of "is-a" is class inclusion relation, but sentence S1 and its paraphrased propositions P1 have a different sense. Proposition P1 asserts the class-instance relationship between 'each mitochondrion' and 'membrane-bound structure'. Therefore, it requires re-representation as shown in RP1, wherein "is-a" is replaced with a specific linking word "is an instance of", which is semantically defined (Smith et al., 2005) as:

> is instance of =def. An object is an instance of a class if it is a member of the set denoted by that class. One would normally state the fact that individual i is an instance of class C with the relational form (C i), but it is equivalent to say (INSTANCE-OF i C)

Using this definition, a specific object is a member of an object of a class. A thing, or a realizable object, is an instance of its class. For example, "Jerry is a mouse", which means that "Jerry is an instance of mouse". Similarly, the proposition, "Mumbai is a city," also means that "Mumbai is an instance of the city". This way, once we know the intended meaning, we replace it with well-defined linking words. In the case of biology text, the objects mentioned with 'each,' 'this,' 'a', 'an', or 'the' describe specific objects which are instantiated. The sentence S1 mentions specific 'each mitochondrion' within a class of a 'double layer membrane-bounded structure'. Therefore, the linking word for re-representation would be "instance of" in place of "is-a".

In Table 3.3, the P1 is re-represented into two propositions, RP1 and RP2. Though both are correct/valid propositions of what is asserted, RP2 is preferred over RP1 without losing the intended meaning and yet being explicit. In continuation, the P2 of S1 has the intended meaning of part, and whole, i.e., the double membrane-bound structure consists of outer and inner membrane. Since this is a meronymy inclusion relation (explained in the following section), the linking word, "with" in P2 is re-represented with "consists of" in RP2. The same P2 can be re-represented in an inverse relation of RP2 as: "Outer and inner membrane part of double layered membrane structure". The richness of the assertion is provided with "consists of", or "has part"/"part of", which are semantically defined. Therefore, the RP2 asserts a part and whole or meronymic inclusion relation. Each of the outer and inner membranes exists within the double-layered membrane structure.

"Has/Have" Linking Words

Typically, the "has/have" linking word is used to convey the part and whole relation, also known as Meronymic inclusion relation, for example, "cell has mitochondria"; "nucleus has nuclear matrix", etc. We illustrate some cases of re-representing the "has/have" linking word as shown in Table 3.4.

The sentence S1 is paraphrased into proposition P1 wherein the linking word "have" is used. The proposition is asserting a part-whole or meronymic relation. The plastids exist as a whole and the DNA and ribosomes exist as parts with respect to this context. To make the part-whole assertion explicit, the P1 is re-represented as RP1 with the linking word "consists of" or "has part" or its inverse relation "part of", which is semantically defined as:

> 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. part of =def. For all c, t, if Cct then there is some c1 such that C1c1t and c part_of c1 at t.
>
> part of: A core relation that holds between a part and its whole

The "part of" definition asserts that c is an instance of C that exists in at a given time, there exists c1 an instance of C1 at the same time, and c1 is part of c. It also asserts that c does not exist except as part of c1 (Markowitz et al., 1992b; Smith et al., 2005). For example, "finger is part of hand".

Let us understand the RP1 by applying the definition. Let DNA and ribosomes be the class C, and let plastids be the class C1. The relation "part of" can be asserted between the classes C and C1 such that it holds for their instances as well, respectively. The DNA and ribosomes of plastids are here assumed not to exist independent of the plastids. Wherever this interpretation is valid, we shall use "part of" or its inverse "has part" relations. One could also use "consists of" linking word as a synonym for "has part".

Several species of inclusion relations are recognized by Winston et al. (1987), broadly classified into three types – class inclusion, meronymic inclusion, and spatial inclusion relations. Meronymic inclusion relations are further distinguished into different species as shown in Figure 3.5. Some of the meronymic relations are discussed below, where we will see the need to distinguish "part of" from "composed of", "portion of", and "located in".

A taxonomy of part-whole or meronymic inclusion relations yields six types of meronymic relations: (1) component – integral object; (2) member – collection; (3) portion – mass; (4) stuff – object; (5) feature – activity; and (6) place – area, because part relates to wholes in different ways (Winston et al., 1987).

These are said to differ in three ways: whether the part and whole are: functional,

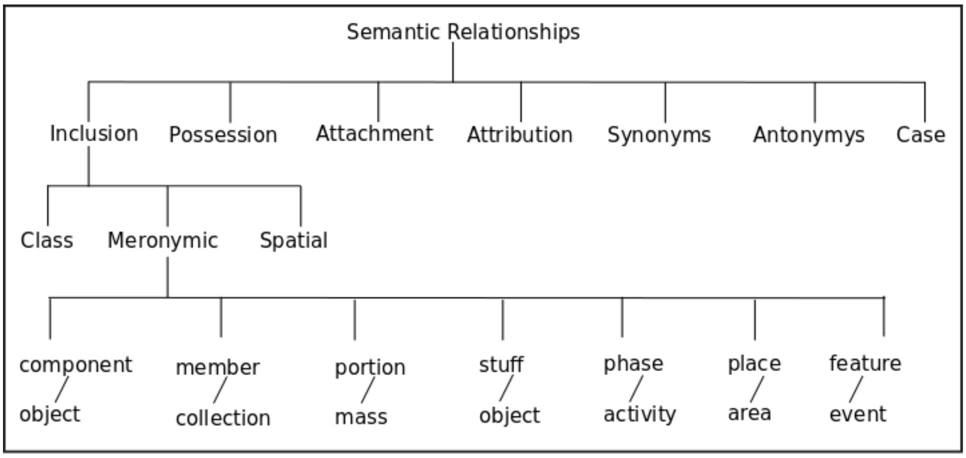


Figure 3.5: Taxonomy of semantic relations (Winston et al., 1987).

homeomerous, or separable. The functional parts in the whole are constrained by their function in space and time (for example, handle – cup). The homeomerous parts are the same as their wholes (for example, slice – pie), while non-homeomerous are different (for example, tree – forest). The separable parts can be separated from the whole (for example, handle – cup), while inseperable cannot (steel – bike) (Winston et al., 1987).

In Table 3.4, the sentence S1, the objects are taking part in component – integral object, the linking word "consists of" / "has part" / "part of" can be used. However, this is not the case with sentence S2, paraphrased as P2. In proposition P2, the subject and object take part in the stuff – object relation. This stuff – object relation can be distinguished from component – object relation as follows. The component-object relation can be elicited by asking, "What are its parts?"; while the stuff-object relation can be elicited by asking, "What is it made of?" i.e., about the composition of the object or constituents of the object. According to Winston et al. (1987), another way to make the distinction between component-object and stuff-object relations is that the stuff with which the object is made cannot be separated, unlike in component-object where the component can be separated from the whole object. Winston et al. (1987) suggests using "made of" (or its synonym "composed of") for stuff – object relations and not "part of". Therefore, the P2 is best re-represented as RP2: "Membrane of erythrocyte composed of 52% proteins and 40% lipids". The definition of "composed of" is discussed below:

composed_of=def. Component parts of anatomy of tissue made up of certain cells or other body area/system or tissue types.

composition=def. A single physical entity inhering in an bearer by virtue of the bearer's quantities or relative ratios of subparts.

We can apply this definition in RP2: A single physical entity (membrane of erythrocyte) inhering (comprising) ratios of subparts (proteins and lipids). Table 3.4 illustrated two sentences S1 and S2 with different relations within the meronymic inclusion relation.

The meronymic inclusion relation is often also confused with the spatial inclusion relation also known as the topological relation (Winston et al., 1987). Let us understand the distinction between meronymy and spatial inclusion relations. The confusion between these two kinds of relations is due to the linking word "has/have", resulting in the ambiguity between meronymic and spatial inclusion relations. We have seen above that the intended meaning of "has/have" is primarily for meronymic or part and whole relation. But the same linking word "has/have" is also used to connote a spatial inclusion relation. There needs to be a distinction between these two kinds of relations. We know that meronymy relations exist between part and whole objects that co-extensively. On the other hand, spatial inclusion relations are described as one object being surrounded by another object but not by being a part of it, which means not co-extensively (not occupying the same space).

Winston et al. (1987, p. 427), clarifies the ambiguity between meronymic and spatial inclusion relations, citing two propositions – P1: "Florida surrounds the Everglades"; P2: "East Germany surrounds West Berlin". Both propositions have the same linking word "surrounds by". In the P1 the Everglades are part of Florida, as they are co-extensive to each other. On the contrary, in the P2, West Berlin and East Germany are not co-extensive, meaning no part of West Berlin is East Germany (Winston et al., 1987). Therefore, P1 is meronymic, but P2 is a spatial inclusion relation and not meronymic.

Let us take a few examples to understand the distinction. In the proposition, "Cell has plasma membrane" or "Plasma membrane is part of cell". This means every part of the plasma membrane is part of the cell and is co-extensive with each other. In this case, meronymy is applied. For further clarity, this proposition can be re-represented as "Plasma membrane (co-extensively) surrounds the cell" (where co-extensively surrounds would mean meronymy).

It is a common usage that the linking word "surrounds" is used in co-extensive and not co-extensive sense. For example, P1: "Cell is surrounded by plasma membrane" and P2: "Nucleus is surrounded by cytoplasm". As explained before, the plasma membrane is not only part of the cell but also is co-extensive with the cell. On the other hand, the cytoplasm is not part of the nucleus and is also not co-extensive. The later usage is more like the "house is surrounded by trees". Though "surrounded by" is used in both cases, this semantic distinction is not brought out by this linking word. To resolve this ambiguity issue, we suggest that whenever there is a case where the objects are co-extensive to each other, we use the linking word "enveloped by" to assert meronymic inclusion by convention. While, the linking word "surrounded by" is to be used for spatial inclusion relations alone, where the objects are not co-extensive to each other. The other option is to qualify, "surrounded by" with co-extensively for asserting meronymy. The linking words, "enveloped by" and "co-extensively surrounded by" can be used synonymously for meronymy.

Let us understand the definitions of these two linking words: "enveloped by" and "surrounded by" as shown below. We propose the definition "enveloped by":

enveloped by: [definition]. X enveloped_by y if and only if (1) y is a part of x, and (2) x is surrounded by y

Applying the above definition, the cell enveloped_by plasma membrane if and only if (1) a plasma membrane is a part of the cell, and (2) a plasma membrane surrounds the cell

The above definition, uses "part of" which is defined above and "surrounded by" which is defined below (Ong et al., 2017):

> Surrounded by = [definition]. X surrounded_by y if and only if (1) x is adjacent to y and for every region r that is adjacent to x, r overlaps y (2) the shared boundary between x and y occupies the majority of the outermost boundary of x

Applying the above definition, the nucleus is surrounded_by cytoplasm if and only if (1) the nucleus is adjacent to the cytoplasm and for every region r that is adjacent to the nucleus, r overlaps cytoplasm (2) the shared boundary between nucleus and cytoplasm occupies the majority of the outermost boundary of the nucleus.

In sentence S3, the SCM proposition, P3 mentions "mitochondrion having two membrane coverings". The P3 asserts a relation of covering or an envelope between mitochondria and membrane. In addition, the mitochondria and the membrane have a part-whole relation. Based on the above analysis and the proposed definition of the linking word "enveloped by", the P3 is re-represented as "Mitochondria enveloped by two membranes" as shown in RP3.

Not only does the Re-representation of linking words make the meaning explicit, but it also helps to resolve the ambiguity. Whenever this situation arises, we suggest consistently replacing such ambiguity with semantically defined linking words.

"Of", "In", "On", "About", "With" Linking Words

If the linking words are *prepositions* (of, with, from, on, etc.) the resulting relations between concepts or terms are expressions and would not be qualified as propositions.

Prepositions such as "of", "in", "on", "with", and "about" are often used as linking words in concept mapping. If these prepositions are combined with a specific linking word, for example "located in" or "contained in" instead of "in" alone, then this can be unambiguous. Prepositions can be used along with other verbs. When these are used without a verb, one may get an expression but not a complete proposition. For example, "Water in a bottle" uses the preposition "in" alone without a verb. But it can be re-represented with, "Water is contained in a bottle", which uses a preposition and the verb "contained in". We suggest re-representing these linking words with semantically defined linking words to bring clarity and specificity to each of these prepositions. Table 3.5 illustrates examples with prepositions as linking words used in sentences from science text.

In Table 3.5, the sentence S1 is paraphrased into propositions P1 and P2. The linking words "lie on", and "buried in" are used in these two propositions. The intended meanings of these two propositions connote the place (location) of the objects within the other objects. According to Winston et al. (1987), these are spatial inclusion relations, wherein the object is located or surrounded but not a part of the object that it locates or covers. For every object, there is some region in which it is located. The objects (entities) that occupy the spatial regions can be denoted with relations, for example "located in", "contained in", and "adjacent to". Examples of these relations, as given in Smith et al. (2005) are: "ribosomes located_in cytoplasm", "lung contained_in thoracic cavity", "inner layer of plasma membrane co-adjacent_to outer layer of plasma membrane".

Although the linking words "lie on" and "located on" are synonymous, for consistency, we will continue to use "located on", as it is semantically defined in the Reference Set, shown in RP1.

We use the linking word "located on" from an existing semantically defined linking word "located in". We propose the definition of "located on" using the definition of "located in" as shown below:

c located_in r at t - [definition]. a primitive relation between a continuant instance, a spatial region which it occupies, and a time proposed: c located_on r at t - [definition]. a primitive relation between a continuant instance, a spatial region on which it occupies, and a time

Another definition by Schulz et al. (2005) states that:

The "location" relation of the spatial inclusion relation has been also formally defined as when each physical object or anatomical entity (c, d) is associated with a spatial region (r)at a time (t) as: located-in(c, d,t) = [definition]. part-of(r(c,t), r(d,t),t)

It means c is located in d at time t, when c is part of the region r at time t, and d is also part of the region.

Continuing with sentence S1 and its paraphrased proposition as P2, the linking word used is "buried in". The assertion is about objects being embedded or immersed inside the other object. We can re-represent the P2 by replacing it with semantically defined linking word based on "immersed in" (Ong et al., 2017) as in RP2:

immersed in: [definition] a relation between a (physical) entity and a fluid substance in which the entity is wholely or substantially surrounded by the substance.

Continuing with the use of prepositions, the sentence S2 is paraphrased to propositions P3, P4, P5. Each of these uses different prepositions, such as "are", "are of", " and "are with" as linking words with varied intended meanings. Propositions P3, P4, and P5 take part in class inclusion, attribution, and spatial inclusion relations, respectively.

The P3 is an example of class inclusion relation discussed above in the section on "is-a"/"are". The P4 has the intended meaning of the varying properties of objects. These relations are categorized as attribution relations (Winston et al., 1987). For example, trees can be tall or short. The property of height is the attribute of trees as objects. The attributes are variables, which means the height of trees can vary. The P4 asserts the varying shapes and sizes of the object leucoplast, and it can be re-represented with the linking word "varies in". We propose the definition of "varies in" borrowing from the semantically defined linking word "completely_varies_in" as shown below. The source is from the portal – http://purl.obolibrary.org/obo/GENO_0000784):

completely_varies_in: [definition]. A relation between two sequences or features that are considered variant with each other along their entire extents.

proposed definition: varies_in: [definition] When X varies_in Y, it implies instances of X can occur in variable instances of Y.

For example, "vacuoles vary in shapes". Applying the proposed definition of linking word, the P4 can be re-represented as "leucoplasts varies in shapes and sizes" as shown in RP3.

In sentence S2, proposition P5 uses the preposition "are with". The intended meaning of the P5 is about containment relation, which is a spatial inclusion relation. It is possible that the P5 itself can be paraphrased by using the linking word "contains" or its inverse relation "contained in" as the word storage connotes the containment of materials in a region. The paraphrased P5 is re-represented as RP4 proposition, using the semantically defined linking word "contained in" (Smith et al., 2005) as shown below:

c contained_in c 1 at t = [definition] c located_in c 1 at t and not c overlap c 1 at t. (Smith et al., 2005)

Another definition is also shown as below:

contained-in(c, d,t) = [definition]. located-in(c, d,t) and not(part-of(c, d,t))

It means, c is contained in d at time t is defined as c located in d and c is not part of d. For example, urine is contained in urinary bladder, means urine is located in urinary bladder and urine is not part of urinary bladder (Smith et al., 2005).

Containment relation is also obtained between a material object (or a substance) and some immaterial cavity or body space (Schulz et al., 2005). By applying the definition, "leucoplasts contains stored nutrients" or its inverse "stored nutrients are contained in leucoplasts" means that stored nutrients are located in leucoplasts without overlapping.

The sentence S3 is paraphrased as proposition P6. The intended meaning is about composition. In the same sentence S3, the P7 belongs to function relation, and its RP7 is the nominalized process-centric proposition. We discuss this in detail in the section on re-representing dynamic proposition (Phase 3).

We now apply the re-representation Phase-1 and re-represent the linking words to create re-represented propositions of the same map as shown in Figure 3.4. While doing so, we ensure that the meaning is preserved and there is no loss of knowledge. If the existing linking words are appropriate, then we preserve the same, and if they appear vague or ambiguous, then we replace them from the Reference Set.

Logical Connectives

Scientific language is just not about describing facts or statements. It also dwells on reasoning, inference, explanation, etc., for scientific modeling. We suggest the use of logical connections to be applied for reasoning between statements. During content analysis, we looked for processes in biology text with the commonly used logical connectives (Copi et al., 1972) such as conjunction (and), disjunction (or), negation (not), material implication (if then), biconditional (if and only if). We find the terms such as, "because", "hence", "due to", "therefore", etc. indicators of logical connectives in the language of science. However, in this work, we do not focus on logical connectives extensively; only mark them in the passages.

Similarly, the other phrase, "therefore," which is a logical connective between two or more propositions, is missed during mapping. Perhaps because the tool has no provision, it gets mapped along with either subject, linking phrase or object. Logical connectives cannot be combined with either of these; therefore, we need to use them independently to connect propositions.

These terms are represented as logical connectives for reasoning between two propositions – "P1 therefore P2"; "P3 because P4". For example, "Leaves have color green" because "leaves contain chlorophyll". More about this is discussed in the section on re-representing propositional linkages (Phase-6).

3.5.3 Preliminary Content Analysis – Growth of Knowledge

We applied Phase 1 of re-representation using well-defined relations for analysis of three textbooks of classes 8, 9, and 11 on the topic of cell structure and function (NCERT, 2007b,c,a). Just as the constrained set of 7 linking words was enough to represent an entire chapter of the topic in the class 8 textbook, we explored the possibility of some more types of linking words that would be required for representing textbooks of higher classes. We got exciting findings. The structure-function based sentences only in the school textbooks were represented using a finite set of linking words (relations). We also observed that these same finite set were repeatedly utilized for all the three levels of school textbooks. We consider this as saturation. However, for the same structure-function based sentences, the linking words depicting properties or attributes showed an increase with the levels of textbooks. Figure 3.6 which shows the list of relations and attributes (Kharatmal and Nagarjuna, 2011). The graph in Figure 3.7 depicts the saturation of relations and increase in attributes. Apart from the structure-function based sentences, there were other forms of sentences which were not able to capture using concept mapping or using the finite set. We call this as residue (also mentioned in section 3.5.1). As list of this residue is – For class 8 - because, called as/known as, cannot, differ from, does not affect, due to, except, exposed to, increase, means, passed, relate to, varies in, while. For class 9 - acts as, affected by, because, becomes, before, called/known as, during, group of, involves process, means, move across, occupy, occurs, related to, required for, resulted in, similar to, undergoes process, visible with. For class 11 - acts as, called, discovered, divides by, duplicates, emerge from, observed by, occupies, proposed by, required for, traversed by, year. This residue was not part of the content analysis in the early explorations study. During this early explorations, this was a limitation of the method as it was able to represent only structure-function based sentences. It is only after the method was further developed for representing processes, we were able to capture most of the residue in later Phases of re-representation.

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

Figure 3.6: 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.

3.6 Re-representing Attributes – Phase-2

We distinguish between relations and attributes while focusing on the linking words. This is significant for representing variable properties of objects (Quine, 1948; Halliday, 2006; Winston et al., 1987). For example, in the sentence, “cisternae are of with many flat, disc-shaped sacs with 0.5 microns to 1.0 microns diameter”, the intended meaning is about a shape and range of diameter of cisternae. So the verb “are of” is ambiguously used for shape and diameter. Since these are variable properties, these are re-represented with the predicate “has shape” and “has diameter range”. Properties enable detailed descriptions of phenomena and quantification in biological processes. Properties depict changes in physical qualities or measurable quantities (Doran and Martin, 2021).

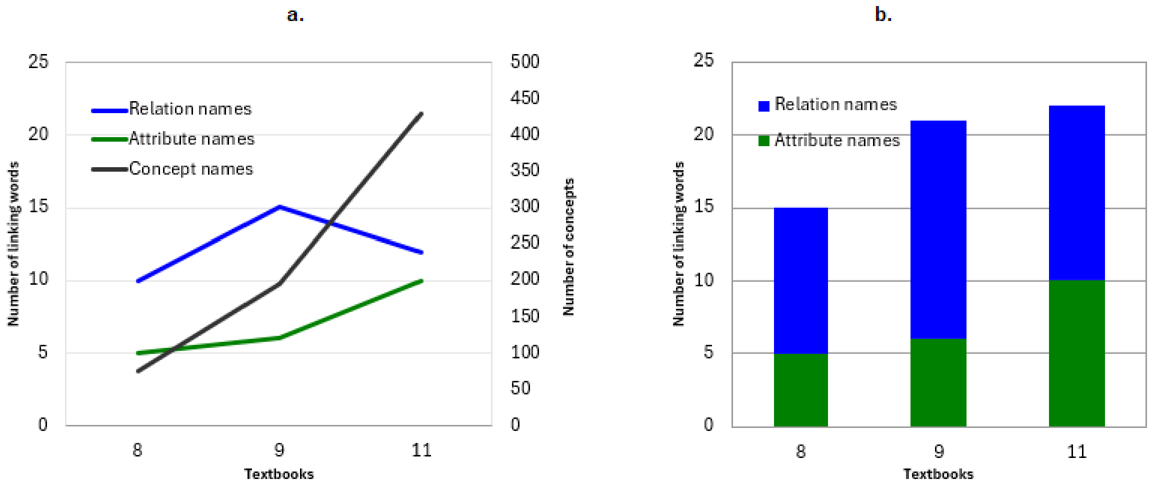


Figure 3.7: a.: Graph depicting constancy in predicate terms even when the concepts increase progressively in class 8, 9, 11. The concept names are scaled on the secondary y-axis (Note: lines connecting points are just a guide to the eye); b.: Graph depicting the proportion of relation names and attribute names linked to the concepts in classes 8, 9, and 11.

Even though these attributes can be represented as linking words, to meet the objective of applying KR principles, these need to be defined distinctly. This is important because it is only due to this distinction that it is possible to represent change and processes further. As a development and addition to the re-representation method in Phase 1, the Phase 2 focuses on the distinction of linking words into relations and attributes.

A relation is between two objects, and attributes are unique relations between class and variable properties. The variables 'range over' a domain of possible values (Quine, 1948). An attribute is a property of an object (Winston et al., 1987; Quine, 1948), that are usually described by adjectives, for example, *color, shapes, size, weights*, etc. Sowa (2003, p.89). In the conceptual spaces framework of representation, Gardenfors (2000) applies the notion of dimension based on perceptions "to assign properties to objects and to specify relations among them" (*emphasis by Gardenfors (2000)*). Based on this framework, Gardenfors (2000) assign dimensions such as *quality, spatial, sensory receptors, and mathematical notions having geometrical structures.* For example, *temperature, weight, brightness, and pitch as quality dimensions; height, width, and depth as spatial dimensions; color, sound, and taste as sensory dimensions; points, planes, lines, and distance as* *mathematical dimensions.*

Following the Knowledge Representation area of research, we can find reference to semantically well defined attributes also called as property value (Barker et al., 2001). For example, *diameter, depth, surface area, fluidity, length, size, etc.* are specifically defined in the context of creating a complete vocabulary for modeling biology textbook.

The biomedical ontology community describes attributes as properties in their classification scheme (Ong et al., 2017). While categorizing terms in the ontology databases, the attributes are classified as — qualitative type (*colour, appearance*, etc.), quantitative type (*number, size*, etc.) (See Appendix).

We use this distinction between a relation and attribution in the Phase 2 of the Re-representation method. Regarding representations, the KR community describes the relations as object properties, and attributes are called data properties (Ong et al., 2017). The World Wide Web Consortium's (W3C) Web Ontology Language (OWL) (OWL, 2012) is a semantic web language designed to represent rich and complex knowledge about things, groups of things, and relations between things (Bechhofer, 2009). OWL's things (entities) are represented as classes, datatypes, object properties, data properties, annotation properties, and Individuals. For our interest and context, the classes are concepts, object properties are relations, data properties are attributes, and datatypes are a range of possible values.

3.6.1 Distinction Between Relations and Attributes

Let us take a simple sentence, "The car is a vehicle". The sentence uses "is-a" linking word and is a class-subclass inclusion relation, where objects in both subject-object places refer to a class of objects. Now let us take the sentence, "The car is red". This sentence also uses a "is-a" linking word but is not a class-subclass inclusion relation. This is because the object car refers not to another class of object but to an adjective or a property of the object, i.e., "red". Properties are variables. Color is the property name (becomes variable), and red is the property value (possible values). Such relations are referred to as "attribution" relations. In another instance of the sentence, "Tree is 6 meters", meaning that the object 'tree' is being described with its height property being 6 meters. In the two sentences, "The paper is made of cellulose" and "The paper is white," the former is a relation, and the latter is an attribution. Similarly, shape as a property is used to describe the paper as an object with the domain of possible values, either square or rectangle.

We make the distinction between a relation and attribution as follows:

- A relation is between two classes or individuals or between a class and an individual.
- An attribution is between classes or individuals and variable property (with possible values).

The properties of objects (i.e., color, height, length, size, etc.) are variable and can have a range of possible values: physical/qualitative or measurable/quantitative. A few more examples of attributes as variable property (domain) and the range of possible values are denoted in Table 3.6.

We introduced this distinction in the textbook analysis of the school level. We found that although there was a saturation of relations, the attributes increased from lower to higher levels of school textbooks shown in Figure 3.7 (Kharatmal and Nagarjuna, 2011). This is of significance. We wanted to extend the study by adding two higher levels, i.e., college levels textbooks.

As part of re-representation, we assign variable properties as attribution. An illustration of re-representing attributes from a passage of Class 11 textbook (NCERT, 2007a) is shown in Table 3.7.

> Typically, mitochondria are sausage-shaped or cylindrical having a diameter of 0.2-1.0 micrometer (average 0.5 micrometer) and length 1.0-4.1 micrometer.

The passage describes the object of mitochondria in terms of variable properties such as shape, diameter, and length. Therefore, the propositions can be re-represented by using attributes "has shape", "has diameter", and "has length" as shown in Table 3.7.

3.6.2 Disambiguation of Attributes

The rationale for making the distinction is to disambiguate the attributes from class inclusion relation as shown in Table 3.8

The sentence S1 is paraphrased to obtain two propositions, P1 and P2. The linking word in P1 "are" denotes two senses: class inclusion and attribution. We re-represent the P1 to RP1 and RP2 to clarify these two relations. The RP1 uses the linking word "is a" to denote the class inclusion relation, while the RP2 replaces the linking word "are" with "have shape" because the pili are mentioned here in the context of an elongated and tubular shape. In this way, by re-representing the method, it not only creates a distinction between relations and attributes but also enables disambiguation.

It may be noted that not all the paraphrased propositions must be re-represented. For example, proposition P2 has the linking word "are made of", whose usage provides the correct sense. Therefore, proposition P2 is preserved even in the re-represented proposition.

From sentence S2, the paraphrased proposition P3 has "are" as its linking word. While its usage is intended for class-subclass, its re-represented proposition RP3 uses the most often used linking word "IS-A" to provide the explicit meaning of class inclusion. The same proposition describes 'fimbriae' as a type of structure with a small property of length. However, for both these intended meanings, there is one linking word. To make the sense explicit of these two meanings, proposition P3 needs to be re-represented distinctly. The same proposition P3 is re-represented to denote the attribute of length, and therefore, the linking word used in RP4 is chosen as "has length", as shown in Table 3.8.

This re-represented proposition can now be used to assert a correlation between two variables, "(tubular shape of pili) has a role in (providing a larger surface area of membrane)". The variable properties and values create a state-change model of representing processes.

Variable properties are required to be the focus (subject) in scientific language because the relation between them builds scientific knowledge. Relations between variable properties can establish process explanation, reasoning, and correlations, and become experimental parameters. For example, "shape of pili is related to membrane surface area". Scientific laws are descriptions of relations between variable properties (Giere, 1990). For example, "volume of gas is related to temperature of gas". Scientific text usually follows this implicitly. Through this re-representation exercise, we are making the character of scientific language more explicit. Focusing on variable properties can be achieved by re-representing the proposition through nominalization. We apply the nominalization based on the SFL framework for representing dynamic propositions as iteration 3 in Phase-3.

We focus on variable properties for representing scientific knowledge because the processes or changes in scientific phenomena/events are represented using variable properties with a range of possible values. Each process can be represented as a change in the variable properties. During a process, at a given time, the state of an object is nothing but a value of its variable property. The process occurs through a change in the state of an object, also called the state-change model as Phase 4.

Here, we suggest following the following rules/disciplines while representing attributes for rigor (Kharatmal and Nagarjuna, 2010):

- Not to use "is-a" for attributes
- Use of "has" along with an attribute linking word i.e., with a well-defined predicate only. For example, "has size", "has color", etc.

Phase 1 and Phase 2 of the development of the re-representation method were sufficient for representing structure terms or static propositions. However, biological science is not just about structure terms alone. It also has the physiological component, structure-function relations, and process modeling. This was also observed as we analyzed biology textbooks from higher levels. Representing the process-centric statements needed to be complemented even in our designing of the re-representation method. The need for creating dynamic propositions in the concept mapping method was also pointed out by the concept mapping community (Miller and Canas, 2008; Derbentseva et al., 2007). This required iteration - 3 to introduce the process-centric dynamic propositions to incorporate the focus on processes. Improvisations based on iterations fit as per the DBR methodology. The further iterations of Phase 3 of the re-representation method are illustrated in the following sections.

3.7 Re-representing Dynamic Propositions – Phase-3

Traditionally, biology as a subject is about structure and function. Usually, the textbook representation of the process is in the context of the role/function of a structure. This agency of structure creates teleological and anthropocentric biology that can be eliminated by focusing on processes compared to structure. The shift in biology from structure-function towards process-centric language has been the focus in philosophy of biology (Nagel, 1979; Rosenberg, 1985; Woodger, 1938) and is of significance in science education (Kharatmal and Nagarjuna, 2013).

The propositions about physical movement, action, change of state, and causal relations are considered dynamic (Miller and Canas, 2008). The process terms are brought into focus through nominalization by re-representing relationships (verbs) or qualities (adjectives) into things (i.e., nouns) (Halliday, 2006). The processes as verbs are re-represented through nominalization (Halliday and Matthiessen, 2004; Halliday, 2006). Nominalization as expression was created historically to represent a new kind of knowledge during the scientific revolution. As a result of experimentation, a new type of scientific discourse was designed for establishing the cause, events, reasoning, etc. (Halliday, 2006). It is a process of making (re-representing) relationships (verbs) or qualities (adjectives) into things (i.e., nouns). The nominalization of statements is carried out for explicating relations between processes (Halliday, 2006, p. 173), during which the process terms are brought into focus for creating dynamic statements or process-centric statements.

In the biomedical domain, the relation "has-function" is depicted as a primitive relation that relates biological structures with functions (Burek et al., 2006; Chaudhri et al., 2014c), and defined as “a relation from an entity to its functions” (Barker et al., 2001). For example, in the sentence, “pancreatic cells secretes insulin”, the predicate “secretes” when nominalized into process term, is re-represented as “secretion of insulin”, giving rise to the proposition as “pancreatic cells have a role in the secretion of insulin”. To highlight the process concepts, we use predicates such as “role of”/ “function of,” “site of”, “has subprocess”, “causes”, etc. This is similar to the “momenting” of dynamical activities for scientific explanations in the SFL framework (Doran and Martin, 2021; Unsworth, 2020). This enables a static proposition to become dynamic or process-centric, an essential aspect of understanding biological processes.

Table 3.10 illustrates a few propositions depicting the re-representation of linking words into process terms. The linking words are usually verbs, and the structure or process terms are nouns. While creating dynamic propositions, the verbs are changed to nouns, resulting in process terms and introducing ‘has a role in’ as a verb. The process by which this re-representation occurs, especially for representing process terms, is called nominalization (Halliday, 2006), a process in which verbs become nouns.

3.7.1 Nominalization of Processes

The verbatim use of linking words (verb-based) in textbooks, such as “produces”, “supports”, “stores”, “secretes”, etc., gives an everyday sense to scientific propositions. This colloquial sense is transformed into a much-needed biological sense by re-representing linking words into concept terms, such as “production”, “secretion”, “storage”, etc. It, hence, becomes more process-oriented biological propositions (dynamic propositions). Such an essential process of transforming from colloquial to scientific proposition is due to re-representation by nominalization (Halliday, 2006).

In Table 3.10, the verbatim linking words (verbs) used in the text are depicted. These are indeed depicted as action verbs because an agent (the object) performs specific biological processes (Rosenberg, 1985). We argue that the essence of these propositions is process-centric rather than structure-based objects. To make these static propositions into dynamic propositions (Miller and Canas, 2008; Kharatmal and Nagarjuna, 2013), so that the processes come into focus, we need to nominalize these action verbs and re-represent them as processes as listed in the re-represented propositions. We can notice that the static objects have now been re-represented as processes, for example, “release of energy”, “regulation of protein traffic”, etc.

The non-nominalized linking words, when nominalized, make the text process-centric by adding processes to the content or bringing process-centric words to the foreground, which otherwise were implicit or in the background. For example, in the sentence, “mitochondria releases energy”, the linking word used is “releases” when re-represented is nominalized into the process as “release of energy”. Similarly, in the sentence, “endomembrane system regulates protein traffic”, the linking word “regulates” when re-represented is nominalized into the process as “regulation of protein traffic”. These nominalized process terms bring the focus of the objects from “energy”, and “protein traffic”, into processes such as “release of energy”, and “regulation of protein traffic” and create dynamic propositions such that the representation is to understand the “regulation of protein traffic” as a biological process vis-a-vis mere “structure of protein”. This enables us to focus on the question of ‘how’ in understanding any phenomenon.

Bringing the focus on processes is an advantage of the re-representation method. With just an additional re-arrangement (Juznic, 2012), the process-based terms now in object place can be rearranged to subject place. While doing this, we need to rearrange the nominalized linking word into an inverse form so that the meaning is not changed; for example, “has function” and “has role” are transformed into their inverse forms as “is a function of”, “is the role of”, respectively. The resulting dynamic proposition with the process term in the subject place is rendered as “release of energy is the role of mitochondria” and “regulation of protein traffic is the role of endomembrane system”. As shown in Table 3.11, all the re-represented propositions with nominalized linking words begin with processes.

3.7.2 Representing Change

We proposed this work in an exploratory study on representing change (Kharatmal and Nagarjuna, 2013). Once the re-representation creates the process-centric dynamic propositions, it would be feasible for process modeling, which is proposed as an advancement in the concept mapping methodology and even in the re-representation method. We do not stop in the development of the re-representation method to enable a dynamic proposition. We leap forward to explicitly design the method for process modeling, thus developing an iteration - 4 over the method.

The iterations so far have been about representing propositions with re-represented linking words (iteration 1), making distinction between relations and attributes (iteration 2). This was followed by the need to represent dynamic propositions with process-centric linking words and hence iteration 3. Now that we can re-represent dynamic propositions with process terms getting in the foreground, how do we represent the dynamics of the process itself? For example, in the process of boiling water, we know that the water is an object undergoing the process of boiling. During this process, several changes occur, such as the temperature of the water changes from cold to hot, and the state of water can change from liquid to steam, depending on the time, heat, etc. The boiling dynamics focus on variable properties such as temperature, state of water, heat, time, etc. As there was a need to represent change, the next challenge was to incorporate these dynamics, changes occurring in terms of variable properties (ie, attributes), into the re-representation method so that these become explicit.

We now discuss yet another iteration - 4 based on iterations 2 and 3 as precursors. Only when we make a distinction between relations and attributes we can utilize the attributes as variable properties to depict change for process modeling and we get processes to foreground by nominalization.

3.8 Re-representing Processes as State-Change – Phase-4

A model for a scientific explanation, as put forward by von Wright (1963), about processes is – pTq, where "p" is the initial state of affairs, "q" is the end state of affairs, and "T" is the transformation. The "T" is a process, and the "p" and "q" are states of affairs considered 'features' of the world. The features in von Wright's model are similar to attributes in the KR model. In Sowa's model of process representation, processes are described as an "evolving sequence of states and events" occurring in time and are classified into discrete and continuous (Sowa, 2003). According to Sowa (2003), events, as opposed to processes, are those changes that occur in discrete steps interspersed with inactive states. In ontology modeling, nominalization of predicates is a distinguishing criterion for processes and events (Baratella, 2020; Guarino et al., 2022).

According to the Object Process Methodology (OPM) of process modeling (Dori, 2002), the process is represented as a change in the state of an object. The three entities in OPM – objects, processes, and states - are building blocks. According to OPM, the objects exist, processes transform objects, and states are used to describe the objects. In the OPM, the process of "melting" is depicted as a change in the state of water from "ice" in the prior state to "liquid" in the post state. Similarly, in the process of "freezing," the change is depicted as a change in the state of water from the "liquid" in the prior state to "ice" in the post state.

The modeling of Dori's work is quite similar to that of von Wright's work, depicting processes as changes in states. "Both states are connected by process links for each property whose value is being changed. Each process consists of a change of one parameter" (Zytkow, 1999, p. 318). We can notice a similar design of applying pre and post-representations in the structure-function behavior (SBF) modeling language wherein functions, which are processes, are being modeled using a set of pre-conditions and post-conditions as applied in the domain of engineering (Goel et al., 2009).

The process modeling stated by Dori (2002), is influenced by the model of von Wright (1963), depicting processes as changes in states in terms of variable properties and possible values, and similar to model-based reasoning (Magnani et al., 1999) characterizing science, and scientific explanations. For example, while illustrating Galileo's inclined plane, which is one of the hallmark experiments of modern science, Zytkow (1999) suggests as a principle for representing variable property as states, "to consider one process per each object involved and each parameter that changes its value", further to connect each state to depict process suggests that, "both states are connected by process links for each property whose value is being changed".

Various process and state related specifications are provided with informal and formal definitions (Schlenoff et al., 2000). For example, to represent relations among activities and among activity-occurrences, in terms of temporal, interval, state, prior-state, overlap, between, during, etc. are well-defined. (See Appendix B).

Mere representation of process-centric terms is insufficient to understand the processes as change. To bring this into focus, we need to model processes in terms of state change. A process occurs through a change in the state of an object. Only through change can there be an explanation of processes (Harré, 1970). During a process, at a given time, the state of an object is nothing but a value of its variable property. Each process can be represented as a change in the value of the variable properties.

In cognitive semantics, based on the model of the conceptual space, the researchers argue for analyzing the action space similar to predicate space (Gardenfors, 2000). Even in their work, modeling of action is about changes in properties of objects, which makes it suitable for analysis of the semantics of verbs (Gardenfors and Warglien, 2012). Mapping of action to result is considered to be important in the event category, wherein "the result of an event is modeled as a vector representing the change of properties of the patient before and after the event" (Gardenfors, 2000). In the conceptual spaces framework, a movement of an object's representation is modeled as changes from one position (start point) to another (endpoint) (Gardenfors and Warglien, 2012). According to this framework, the movement in the points (start, end) is considered a vector representing the change of an object's properties, introducing it as kinematics. Further, change can now be defined with three notions: "(1) a state is a point in a conceptual space; (ii) a non-zero vector in such space represents a change of state; (iii) a path is a concatenation of changes of states" (Gardenfors and Warglien, 2012, p. 163).

In the KR domain, the works related to ontological modeling have introduced the ontological framework for systems biology (Podkolodnyy and Podkolodnaya, 2016) and experimental sciences (Hafner and Fridman, 1996; Fridman Noy and Hafner, 2000). For representing experimental materials and methods and the reasoning task, Hafner and Fridman (1996) introduced new ontological categories of *Mixtures* and *Transformations*, along with *made-of* and *sequence* relations. In continuation with the modeling of experimentation, Hoehndorf et al. (2011) included four basic classes for process representation: *Separate*, *Combine*, *Transform*, and *Assess* depending on single or multiple inputs and outputs. A list of binary relations: *part of*, *participates in*, *has input*, *has output*, *modifier of*, *function of*, *realizes*, *occurs in*, *quality of*, were used to model process ontology for integrating systems biology with biomedical ontologies (Hoehndorf et al., 2011). A list of relations specific to representing states: *subevent*, *rawmaterial*, *first event*, *next event*, *object*, *result*, *has state*, etc. were used for depicting change (Chaudhri and Inclezan, 2015). Our work applies these relations to modeling the processes as changes in the states of objects. In all of these frameworks for process modeling, we find some common features, such as variable property describing the state, values of variables, and the change in state as a change in the variable's value.

According to the Systemic Functional Linguistics (SFL) approach, biological processes are dynamic activities that describe scientific phenomena or explanations with a series of events, called 'momenting' (Doran and Martin, 2021; Halliday, 1985). Each explanation is represented by *activity, composition, classification, and property.* This complements our re-representation methodology of *process, meronymic inclusion, class inclusion, variable properties.* The SFL group is creating a compositional characterization of the momented activities; for example, 'prophase' is 'momented' as 'chromatin condensation' and 'chromatin folding.' The SFL group acknowledges the property for a more detailed description and quantification of processes. We are extending the SFL's approach and modeling the processes at a deeper level beyond just compositions or classifications by proposing an additional depth that the processes can be re-represented as prior and post states (Dori, 2002; Sowa, 2006; von Wright, 1963). This proposal has been accepted by the SFL community (Kharatmal, 2023a).

Thus, we develop Phase 4 of representing processes in terms of changes in states of objects.

3.8.1 Process Modelling

We illustrate process representation with a content sample from a widely used as textbook (De Robertis and De Robertis, 1985, p. 420).

Content Sample - 2: Mitosis Cell Division – Prophase

> "The beginning of prophase is indicated by the appearance of the chromosomes as thin threads inside the nucleus. The word "mitosis" (Gr., mitos, thread) expresses this phenomenon, which becomes more evident as the chromosomes condense. The condensation occurs by a process of folding of the chromatin fibers. At the same time, the cell becomes spheroid, more refractile, and viscous".

In this passage, the processes mentioned are "condensation", "folding of chromatin fibers", "changes in cell". The mere mention of processes would not provide explanation about what actually happens to chromatin during chromatin condensation, the folding of chromatin fibers. Similarly, if there are specific changes in the cell, then one cannot know the prior state as well as the post state of the cell. If a process changes an object, it certainly requires that change be represented. We suggest representing these changes using variable properties, such as foldings, length, shape, viscosity, etc. This is also one of the main reasons for distinction between relation and attribution. The previous iteration 2 in Phase-2 serves as a pre-cursor for this Phase-4.

The process of chromatin condensation occurs through two sub-processes: "chromatin folding" and "shortening of chromatid". During chromatin folding, the object chromatin changes concerning its length, which is a variable property. The process is depicted using changes in the state of the object i.e., of chromatin, with prior state and post state. The prior state is – chromatin is less folded, longer in length, and its post state is – chromatin is more/densely folded, shorter in length. Both the states are connected by process-centric linking word – becomes/changes into. Simultaneously with this process, the cell also becomes spheroid, more refractile, and more viscous. From the above passage, we can identify the objects, processes, and variable properties, as shown in Table 3.12.

For representing the prior state and post state in the form of propositions, applying the node-link-node format (subject-predicate-object) form, the folding of chromatin fibers can be re-represented as follows:

- Chromatin fibers have the appearance of loosely folded
- Chromatin fibers have the appearance of densely folded
- Chromatin fibers have a length longer
- Chromatin fibers has length shorter

This representation provides the changes in terms of prior-state and post-state. However, in order to resolve the redundancy of some terms, we can re-represent the attributes that are in object place into subject place by nominalization (Halliday, 2006) shown as:

- Loosely folded Chromatin fibers
- Densely folded Chromatin fibers
- Longer chromatin fibers
- Shorter chromatin fibers

The following re-representation can now link the change in terms of two prior state and post state of the process folding of chromatin fibres. Since we are now linking two propositions specific to state change, we suggest using process-centric linking words. For example, in this case, the linking word "becomes" is used to link two propositions with variable properties, as shown below:

- Loosely folded chromatin fibres become densely folded
- Longer chromatin fibres become shorter OR
- Length of chromatin fibers becomes shorter

With these series of re-representations, the process of chromatin folding can be shown as state change. We apply the same method and represent the other processes as state change as shown in Table 3.13.

3.8.2 Nominalization of Processes

According to the Systemic Functional Linguistics (SFL) framework, nominalization falls under the grammatical metaphor idea whereby nouns nominalize verb processes without changing the meaning of the text. Nominalization is discussed with illustration in chapter 2, section 2.4.

In continuation, re-representing the other processes from the content sample-2: Mitotic Cell Division–Prophase:

A proposition in the form of depicting attributes as a precursor for a state change.

- Nucleoli has size larger
- Nucleoli has size smaller

Linking of two propositions with a process-centric linking word:

- (Nucleoli has size larger) becomes (Nucleoli has size smaller

Changing the attribute to subject place by nominalization:

- (larger size of nucleoli) becomes (smaller size of nucleoli)

Eliminating the redundancy

- larger size of nucleoli becomes smaller

- size of nucleoli becomes smaller

Here, we agree that even if we started with making the relations and attributes explicit for depicting change as prior-state and post-state, eventually we can notice that the relations and attributes are getting implicit, and the re-representation of statements is more of a condensed or concise form. Then the important question is: Why to do this re-representation? For this our justification is that while doing this series of re-representations, we are actually mapping 'changes' and these changes are made explicit while representing scientific phenomena or scientific explanation of processes. Therefore, we think that this re-representation to depict change is of importance to think in terms of 'how' the changes occur in processes or phenomena (Kharatmal and Nagarjuna, 2013).

We illustrate the graphical representation using CmapTool's nested nodes method to depict the process modeling of state change as shown in Figure 3.8.

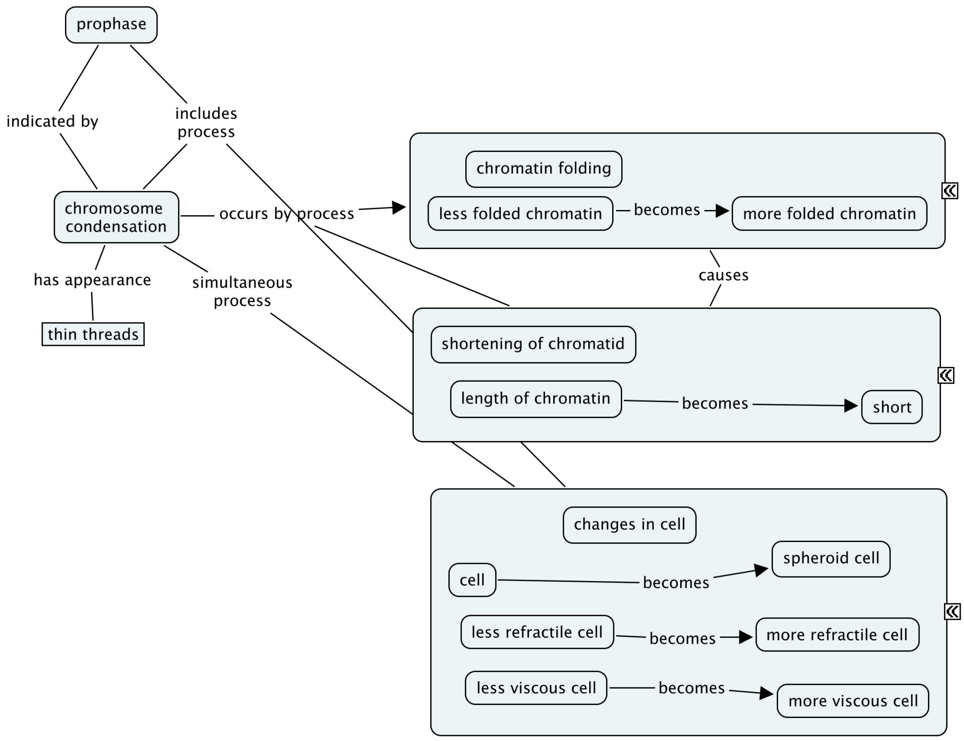


Figure 3.8: Concept map depicting process modelling of prophase of mitotic cell division.

3.8.3 A Framework for Process Modelling

We briefly describe some terms related to process representation as below:

Object/Structure term: An object or a changing system, for example, larva, plant, cell, etc.

Process term: A term that depicts any change of an object, for example, boiling, raining, respiration, pumping of blood, growth, etc.

State of an object: We consider the state of an object as the attributes (properties) of an object at a given time. For instance, the properties of plants can be height, color, etc.

Prior-state and post-state: As the terms depicts, prior-state describes the state of the object before undergoing a process, and post-state describes the state of the object after undergoing a process. In the process of "freezing of water", the prior state of water is liquid, whereas the post-state of water is ice. For example in the process of "chromatin condensation", the prior state of chromatin is longer length, whereas the post-state of chromatin is shorter length.

Linking words for linking states and process: The recommended linking words to link two states can be – "becomes", "changes to" ("transforms into", "converts into"), "moves to", whereas the linking words suggested for linking two or more processes can be – "has sub-process", "co-occurs", "followed by", "causes"/"results in", etc.

A process term is linked with either an object or a process term. If an object term is linked to a process term, we use the linking word – "has a role in". If a process term is linked to a process term, we use an appropriate linking word, e.g., "has sub-process". If several processes occur at the same time, then we could use the linking word – "co-occurs". When the sequential or cyclical processes are involved, we use linking word – "followed by". Another prominent form of linking word – "causes" ("has effect"/"results in") is used to connect the causal relations in the process. As in most of the processes, the change that occurs in the form of state-transformations or state-changes, we depict these with the linking words – "becomes", and wherever applicable, if the process changes the location of the object, it is depicted with the linking word – "changes to", etc. The list of process-based linkages (linking words) is shown in Table 3.14.

By using CMapTool, Figure 3.9 is about the cyclical process, and Figure 3.10 is about cause and effect.

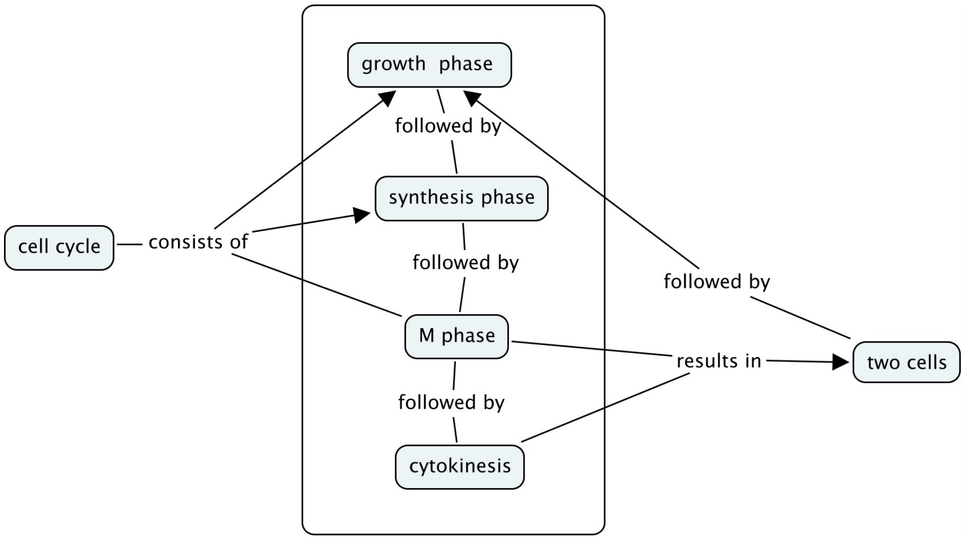


Figure 3.9: Concept map depicting process modelling of cyclical process.

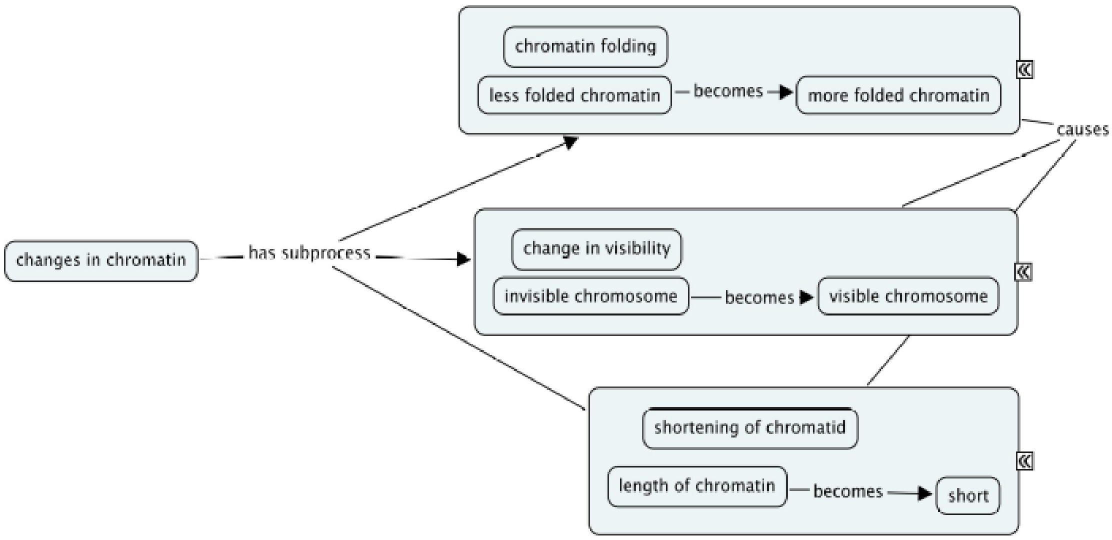


Figure 3.10: Concept map depicting process modelling of cause and effect.

We find parallels in the works of Chaudhri et al. (2014a) that an explicit representation of knowledge is important for creating a knowledge base. This work is using terms from upper ontology to map the objects for explanation. In our method, explanation of scientific phenomenon can be in terms of change in state of objects. Taking an example of the process of signal reception (shown as concept graph), the process Attach is depicted using the linguistic terms form upper ontology such as base object, event, sub-event, next-event, result, etc. Using our method, we could represent the same example as below:

The two processes are about binding and changing:

(i) In process of binding (of signaling molecule to receptor protein):

signaling molecule has position base molecule (prior state)

signaling molecule has position receptor protein (post state)

signaling molecule located on cell's surface Or

signaling molecule located inside cell

(ii) In process of changing the form of receptor protein:

receptor protein has prior state receptor protein1 (prior state)

receptor protein has post state receptor protein2 (post state)

(iii) (binding, changing) initiates (transduction)

By using logical connectives, our method also can be used for reasoning between statements. Although in this thesis we have not showcased it more prominently, but there is a possibility.

To summarize, we adopted the Dori (2002) Object Process Modelling approach and Halliday (2006) nominalization approach for developing the method for process representation as a change in states of objects. Figure 3.11 shows a process-centric framework for mapping processes applying the design principles.

We proposed this model of process representation depicting as a state change model. This was accepted by the concept mapping community and science education research community as well (Kharatmal and Nagarjuna, 2013). As there is scope to utilize the nested node utility of CmapTool, we suggest this as a model to incorporate processes in studies by the concept mapping research work.

3.9 Compiled List of Categories and Rules

The list of semantic categories shown in Table 3.15 is not exhaustive, even for a given domain. This serves the limited purpose of demonstrating the feasibility of the relational content analysis suggested. One may extend or modify the rules and the corresponding

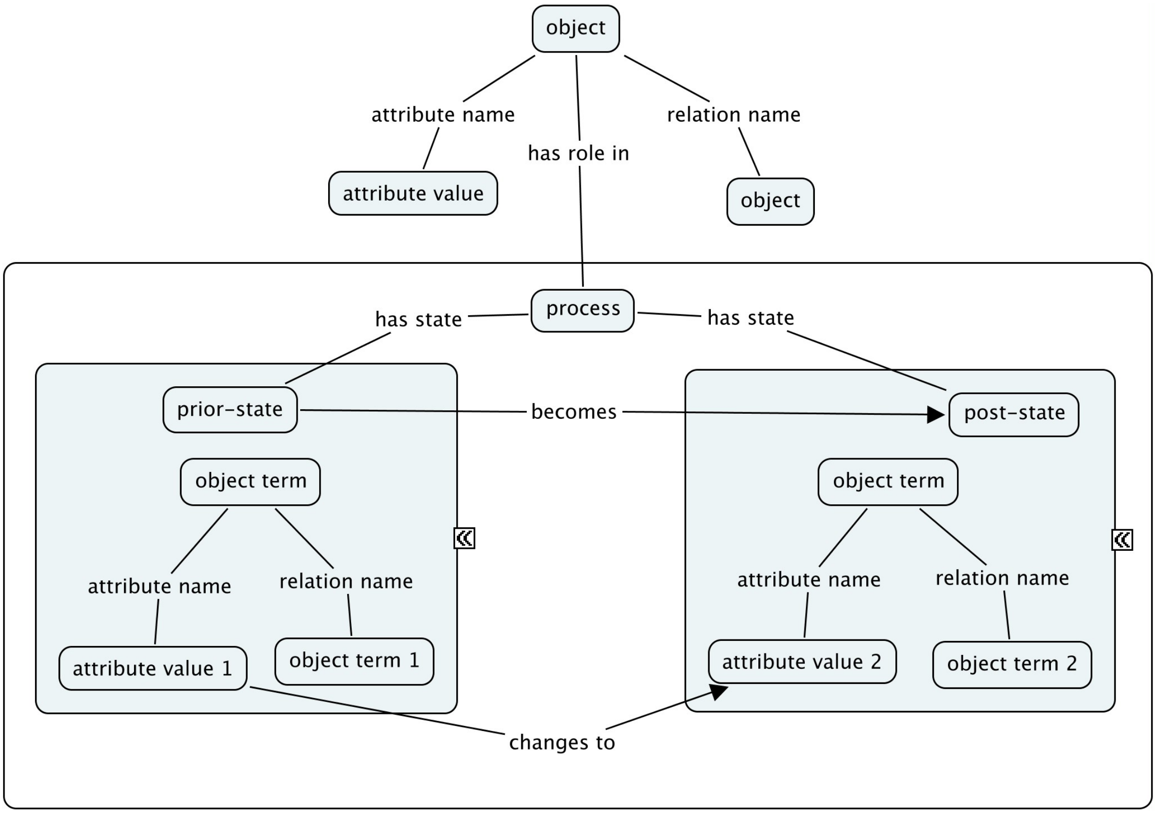


Figure 3.11: Concept map depicting a framework for process modelling.

codes if the idea finds any value.

3.9.1 Semantic Categories

The semantic categories are shown in Table 3.15.

3.9.2 Re-representation Rules

The re-representation rules are shown in Table 3.16.

3.10 Illustrations

We now provide illustrations of some passages applying the semantic categories and re-representation rules. (Also see Appendix for more such passages).

(1) Typically [mitochondrion] is sausage-shaped or cylindrical having a diameter of

0.2 - 1.0 micrometer (average 0.5 micrometer) and length 1.0 - 4.1 micrometer.

In this passage, the intended meaning is about describing the shape, diameter, and length as data properties. The verb is is replaced with has shape and the verb having is replaced with has a diameter of, has a length of and coded accordingly. By applying the semantic categories and following the re-representation rules, the same passage is re-writen as:

(1′) Mitochondrion has sausage or cylindrical shape, [R2, C7] has a diameter of [R3, C8] of 0.2 - 1.0 micrometer (average 0.5 micronmeter) and has length of [R3, C8] 1.0 - 4.1 micrometer.

(2) Plastids also have their own DNA and ribosomes.

In the sentence (2), the verb have is used to for expressing the semantic relation: consists of, belonging to meronymic inclusion. The relations asserts one being a *part* of another. Therefore it is re-written as:

(2′) Plastids also consist of [R3, C1] their own DNA and ribosomes.

In all the above examples, we largely considered sentences that describe structures. Biological text largely includes a description of structure and function. Biology text also has passages that describe processes. We shall perform a relational content analysis of process-centric sentences.

3. Cell membranes separates contents of cells from their external environments, controlling exchange of materials such as nutrients and waste products between the two. They also enable separate compartments to be formed inside cells in which specialized metabolic processes such as photosynthesis and aerobic respiration can take place. Chemical reac-

tions, such as light reactions of photosynthesis in chloroplasts, sometimes take place on the surface of the membranes themselves.

The passage is re-written as:

(3′) Cell membranes have a role in [R6, R7, C5] the separation of contents of a cell from their external environments, the formation of separate compartments inside a cell, the control of material exchange, and [R10] nutrients and waste products between [R1], the two. Special metabolic processes includes [R2, C2], photosynthesis and aerobic respiration, which occur in [R1, C4] the separate compartments of cells. Chemical reactions includes [R2, C2] light reactions, occur in [R1, C4] chloroplasts during photosynthesis, which occur on [R1, C4] surface of membranes.

Passage (3), describes the function of the cell membrane. Therefore, we replace the verb forms by nominalization (Halliday, 2006), where separates becomes "separation of content", forming becomes "formation of", and controlling becomes "control of" by bringing them to the subject place in the sentence. These nominalized process terms occur as class terms in ontologies (Ong et al., 2017). In the nominalized statement, the verbs are replaced by a well defined predicate "has role/has function" (Arp and Smith, 2008b) available from the Reference Set. The re-written statements depict the processes at object place (although these can be at either object or subject place) Further, with just an additional re-arrangement (Juznic, 2012), these process terms can be brought to focus in subject place, using the inverse name of the object property, while preserving the meaning. For example, "The separation of contents of cell is the role of a cell membrane".

The term such as has the sense of subtyping, therefore we replace it with defined object property includes (an inverse name of object property is). The verbs takes place in depict the sense of site or location. Although it's expression is correct, we use its synonym occur in as specified in the Reference Set. One may also note that the prepositions in and on indicate locative prepositions (Zwarts and Gärdenfors, 2016) and following these have been rewritten in combination with explicit predicate terms. The conjunction and is used to combine the two subjects. The relation between is a well defined three place predicate with three arguments (Sowa, 1984), and therefore remains unchanged.

Scientists approach phenomena/processes not merely by naming the terms but modeling by identifying the various parameters/variables, where the change is modeled in terms of change in state.

Consider the following passage:

4. The beginning of prophase is indicated by the appearance of the chromosomes as thin threads inside the nucleus. In fact, the word "mitosis" (Gr., mitos, thread) is an expression of this phenomenon, which becomes more evident as the chromosomes start to condense. The condensation occurs by a process of folding of the chromatin fibers. At the same time, the cell becomes spheroid, more refractile, and viscous. (De Robertis and De Robertis, 1985, p. 420).

The passage is re-written as:

(4′) Chromatin condensation begins in [R8, C6] prophase and ends in [R8, C6] metaphase. During condensation the unfolded chromatin becomes [R8, C6, C7] folded. The thin threads of chromosomes becomes condensed chromosomes. During the process, the cell becomes spheroid, more refractile and viscous. The term “mitosis” refers to [R1] the thread formation, which is derived from [R1] “mitos” meaning thread.

The passage refers to several state changes. To retain the temporal order, we modeled the process as prior state and post state as a requirement of process modeling elaborated in the previous section. We use becomes as a semantic relation between prior state and post state. “The unfolded chromatin becomes folded” or “thin threads becomes condensed, thick chromosomes”. Similarly, refers to is a defined object property for naming. Two or more class terms or adjectives can be combined with the usage of prepositions to create condensed expressions. This passage represents processes as a state change model, described in the section 3.8.

In this section, we presented some illustrations of relational content analysis of science text focusing on object properties (relations) and data properties (attributes). The emphasis is on rewriting the sentences to arrive at an explicit form. This re-statement of a sentence is categorized as *explicitly stated knowledge* in KR (Chaudhri et al., 2011). As part of the re-representation of relations, our approach has been towards resolving ambiguity, making implicit meaning explicit, making distinctions between object and data property, nominalization of verbs for processes, and representing processes in terms of state change. We followed the semantic codings and applied the re-representation rules as mentioned in section 3.9.

The Phases 1-4 shall be used for semantic content analysis (chapter 4, chapter 5 and for evaluation studies (chapter 6). We re-iterate that the following Phases 5 and 6 serve as an extension of the method for illustrating representation of experimentation. In Phase 5 we characterize experiments by analyzing students' representations, and in Phase 6 we represent an experiment by applying procedures following the previous Phases. These two Phases 5 and 6 are also not part of semantic content analysis, student evaluation, feasibility/efficacy studies, proximity studies.

3.11 Re-representing Experimentation – Phase-5

As an iteration and further extension of our method, we characterize experiments following the SFL framework. Scientific explanations are characterized by nominalization, a higher percentage of action verbs organized in a logical sequence, whereas experiments are characterized by the use of imperatives (Halliday and Martin, 1993).

3.11.1 Multimodal Analysis

By identifying the variable properties and parameters (a standard way of experimental design), we show that the proposed content analysis methodology also applies to passages describing experiments. In addition to the textbook analysis, we are doing a multimodal content analysis (Martin et al., 2021; Unsworth, 2020) of experimental design following SFL. The undergraduate biology students' description of experimental design about olfactory responses of Drosophila larvae is published in (Kharatmal et al., 2022). A sample of the description and drawing is shown in Figure 3.12.

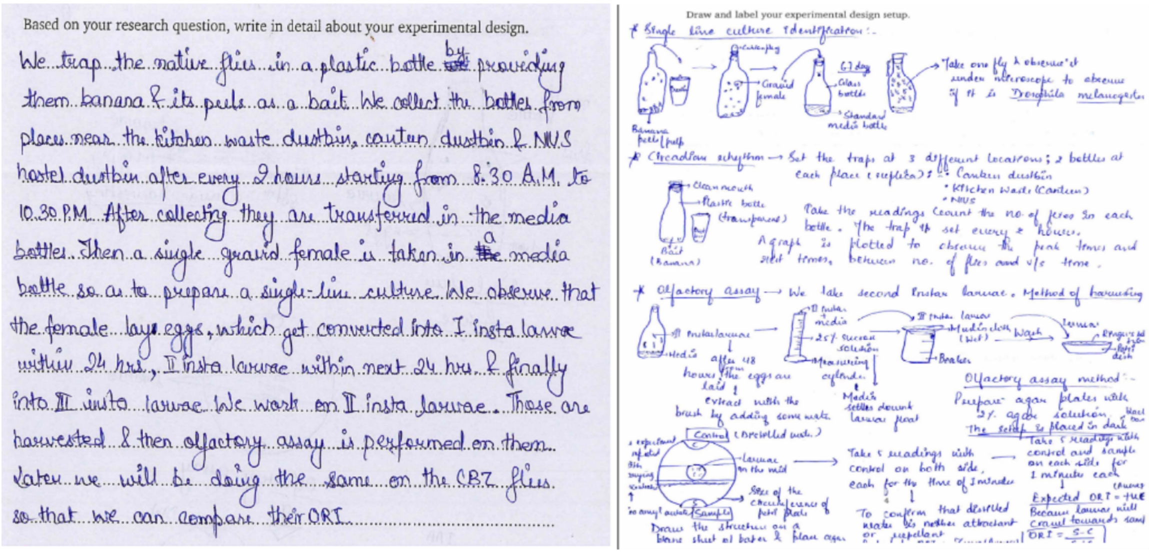


Figure 3.12: Undergraduate student's descriptions and drawings of experimentation.

In the analysis of students' descriptions, a progression is observed comprising of compositional relations interconnected with activity. The intermodal construal of composition and activity was observed in image and language for at least two activities. The 'trapping of fruit flies' comprises of keeping banana bait in a bottle, and observing for trapped flies. The 'setting up of single line culture' includes the setting of preparation of media, gravid female fruitflies, selection of fruitflies based on external morphology, and transferring to new media. This was depicted as a linear or cyclical sequence with at least 3-4 micrographs or standalone annotated images. The 'recording ORI' activity was depicted using images and verbiage with annotations. Students have also described the variable properties while depicting the experimental design, such as 'duration' of eggs hatching into larvae, 'percentage' of agar media and sucrose solution, 'recording time' of observation, 'concentration' of dilution, 'measurement' of Ringer's solution, etc. The SFL considers these as just property. However, we extend these properties to quantify their values or attributes as prior and post-states to depict change. For example, the 'change in concentration of dilution' from 10-1 (prior-state) to 10-2 (post-state) would result in a change in the ORI (olfactory response index) of larvae. These findings are reported to the SFL community and have been accepted (Kharatmal, 2023b).

We argue that, though the experiment genre may have commonalities with the explanation genre, but addition to it, using the variable property to depict the rich descriptions and quantification seems distinctive in the experiment genre with the use of imperatives. Our proposal for the re-representation of experimental design adds and extends the SFL framework.

The relationship between experimental and model-based reasoning is already well-established in the literature (Magnani et al., 1999; Zukswert et al., 2019). These investigations of students' multimodal representations of experimentation can bear pedagogical implications in biology education.

3.12 Re-representing Propositional Linkages – Phase-6

We have designed the method of re-representation of relations, attribution, processes (explanations), and experiments. But scientific language is not just about describing facts or statements. It also dwells on reasoning, inference, etc., for scientific modeling. We propose the re-representation of propositions for reasoning between statements. We analyze the Hershey and Chase experiment from text passages (NCERT, 2007d) and apply the various phases of re-representation for a scientific explanation of processes, for reasoning and making inferences, using logical connectives (Copi et al., 1972).

While designing this specific part of the re-representation methodology, we looked for processes in biology text with commonly used logical connectives such as conjunction (and), disjunction (or), negation (not), material implication (if then), biconditional (if and only if). We also looked for text with indicators that connect two or more statements that provide reasoning, such as, "because", "hence", "due to", "therefore", etc.

According to Halliday and Martin (1993), logical connectives are not just a supplementary text but are an essential part of science text because these are used for explanation, reasoning, and justification in science (Evagorou and Osborne, 2010).

We now explore the re-representation method so that it can be shown to be applied for reasoning, making inferences, explanation, scientific modeling, etc. We could identify various ways of reasoning (inductive, deductive, analogical, model-based) in statements. The conclusion statements are deduced in an ordered sequence of premises. The rules of inference are used to build reasoning. Any standard textbook on propositional logic, e.g.,Copi et al. (1972) has laid down these rules. These rules of inference state that the inference is valid if the conclusion follows from a set of premises.

3.12.1 Analysis of an Experiment

We deconstructed the passage into a series of individual statements or propositions. These were then linked with other propositions using connectors. We illustrate one section of Class 12 text on the topic of "Molecular Basis of Inheritance" – Hershey and Chase experiment (NCERT, 2007d, p. 101-102). We specifically chose this section because of its significance in biological understanding. It has an experimental model, fact-based statements, observations, and conclusion statements. As this experiment is an excellent example of scientific method modeling, its re-representation would also be of significance or value.

Content Sample - 3: The Genetic Material is DNA

The source is from NCERT (2007d, p. 101-102).

The unequivocal proof that DNA is the genetic material came from the experiments of Alfred Hershey and Martha Chase (1952). They worked with viruses that infect bacteria called bacteriophages.

The bacteriophage attaches to the bacteria and its genetic material then enters the bacterial cell. The bacterial cell treats the viral genetic material as if it was its own and subsequently manufactures more virus particles. Hershey and Chase worked to discover whether it was protein or DNA from the viruses that entered the bacteria.

They grew some viruses on a medium that contained radioactive phosphorus and some others on medium that contained radioactive sulfur. Viruses grown in the presence of radioactive phosphorus contained radioactive DNA but not radioactive protein because DNA contains phosphorus but protein does not. Similarly, viruses grown on radioactive sulfur contained radioactive protein but not radioactive DNA because DNA does not contain sulfur.

Radioactive phages were allowed to attach to E. coli bacteria. Then, as the infection proceeded, the viral coats were removed from the bacteria by agitating them in a blender. The virus particles were separated from the bacteria by spinning them in a centrifuge.

Bacteria which was infected with viruses that had radioactive DNA were radioactive, indicating that DNA was the material that passed from the virus to the bacteria. Bacteria that were infected with viruses that had radioactive proteins were not radioactive. This indicates that proteins did not enter the bacteria from the viruses. DNA is therefore the genetic material that is passed from virus to bacteria.

We will begin with representing statements or propositions from the passage, assigning them with a letter.

Let us begin with facts-based statements:

- a = (DNA contains phosphorous)
- b = (proteins contains sulphur)

We know that a and b are facts, and we know about it. We can create propositional linkages by using the connector "and" such that:

a and b

(DNA contains phosphorous) and (proteins contain sulphur)

Now we shall represent the experiment based statements:

Experiment-1:

- c = some viruses grown on radioactive phosphorous
- d = some viruses grown on radioactive sulphur

Following are observation based statements:

Observations:

- e = some viruses contained radioactive DNA (Rule of hypothetical syllogism)
- f = some viruses did not contain radioactive DNA (Rule of Disjunctive syllogism)
- g = some viruses contained radioactive protein
- h = some viruses did not contain radioactive protein

Following was the result from the Experiment-1

Results:

(some viruses contained radioactive DNA) because (some viruses grow on radioactive phosphorous) and we know that (DNA contains phosphorous) (Rule of Addition)

(some viruses contained radioactive protein) because (some viruses grow on radioactive sulfur) and we know that (DNA does not contain sulfur) or (protein contains sulfur)

We have used the connectors "because", "and we know that".

The reasoning that has emerged in the results is as follows: an observation statement is linked with experimental statements, and based on the facts, we could provide reasoning for the experiment's results.

In continuation, the following are statements related to Experiment -2:

Experiment-2:

- i = viruses (radioactive phages) attached to E.coli bacteria

The following are observation based statements:

Observations:

- j = E.coli bacteria developed infection
- k = E. coli bacteria were found to be with radioactive DNA
- l = j because i and since e and also k

l = (E.coli bacteria developed infection) because (viruses (radioactive phages) attached to E.coli bacteria) and since (some viruses contained radioactive DNA), and also (E. coli bacteria were found to be with radioactive DNA)

In this propositional linkages, the reasoning is applied by connecting observations, with statements from experiment-1 and experiment-2, and then an inference is made.

Following was the result from the Experiment-2

Results:

- m = DNA transferred from virus to bacteria

The conclusion statement from the above two experiments is as below:

Conclusion:

- n = DNA is genetic material
- l indicating m and, therefore n

(E.coli bacteria developed infection) because (viruses (radioactive phages) attached to E.coli bacteria) and since (some viruses contain radioactive DNA), and (E. coli bacteria contain radioactive DNA) indicating (DNA movement from virus to bacteria) therefore ([it is possible that] DNA is a genetic material)

Here, inference rule is the rule of modus tollens.

In the conclusion statement, all the above set of propositional linkages, (l), the experiment indicates a result and "therefore" as a connector is used to make the conclusion statement.

In the propositions and propositional linkages, there appears to be a pattern, at least for modeling an experiment. Given some facts, an experiment is designed, observations are made, and a conclusion or inference of the results is made using reasoning. Thus, we have illustrated the potential of the re-representation method for propositional linkages that can be applied to analyze statements with reasoning.

3.13 Summary

This chapter contains the core argument of the thesis. We demonstrated a way of performing semantic (relational) content analysis, from simple examples to more complex ones. This study resulted in arriving at a set of codes that can be used for content analysis. Even though the re-representation method is depicted phase-wise, the method can be applied simultaneously, seamlessly interchanging the sequence during re-representation. The phase-wise iterations of the method are:

- Re-representation Phase - 1: Re-representing the relations (linking words) (applying open biomedical ontologies, library of generic concepts)
- Re-representation Phase - 2: Distinction of relations with attributes (applying open biomedical ontologies, library of generic concepts, conceptual spaces)
- Re-representation Phase - 3: Re-representing dynamic propositions (applying nominalization, process specification language, library of generic concepts)
- Re-representation Phase - 4: Re-representing process as change for process modeling (applying Knowledge Representation, Object Process Modeling, process specification language, library of generic concepts, open biomedical ontologies, conceptual spaces, nominalization)

- Re-representation Phase 5: Representing experiments (applying the SFL framework)
- Re-representation Phase 6: Propositional linkages for inference-based statements

We emphasize that this is not an exhaustive iteration. If this method is found valid, we welcome researchers to take the development further depending on the domain of the study. For example, for other topics within biology or physics, chemistry, and other sciences.

With this, we have addressed the research question 1. To address research question 2, we apply this method for semantic content analysis of science text. In chapter 4, we illustrate the use of the re-representation method in relational content analysis for analyzing 6 biology textbooks and students' representations. We also demonstrate the coding and rules followed during the relational content analysis.

No.	List of Sentences (S)	No.	List of Propositions (P)
S1	Mitochondria are known as the powerhouses of the cell	P1	Mitochondria are known as powerhouse of cell
S2	The energy required for various chemical activities needed for life is released by mitochondria in the form of ATP (Adenosine triphopshate) molecules	P2	Energy required for various chemical activities
		P3	Various chemical activities needed for life
		P4	Energy released by Mitochondria or Mitochondria releases Energy
		P5	Energy in the form of ATP molecules
S3	ATP is known as the energy currency of the cell	P6	ATP molecules known as energy currency of cell
S4	The body uses energy stored in ATP for making new chemical compounds and for mechanical work	P7	Body uses Energy or Energy used by body
		P5	Energy is stored in the form of ATP
		P8	ATP molecules used for making new chemical compounds
		P9	ATP molecules used for mechanical work
S5	Mitochondria have two membrane coverings instead of just one	P10	Mitochondria have two membrane coverings (instead of just one)
		P11	Membrane consist of outer membrane
		P12	Membrane consist of inner membrane
S6	The outer membrane is very porous while the inner membrane is deeply folded	P13	The outer membrane is very porous
		P14	Inner membrane is deeply folded
S7	These folds create a large surface area for ATP-generating chemical reactions	P15	Deeply folded creates a large surface area
		P16	Large surface area is used for ATP generating chemical reactions
S8	Mitochondria are strange organelles in the sense that they have their own DNA and ribosomes	P17	Mitochondria are strange organelles
		P18	Mitochondria have their own DNA
		P19	Mitochondria have their own ribosomes
S9	Therefore, mitochondria are able to make some of their own proteins	P20	Mitochondria, therefore can make their own proteins

Table 3.1: List of sentences extracted from the passage and corresponding propositions.

No.	List of Propositions (P)	No.	List of Re-represented Propositions (RP)
P1	Mitochondria are known as power house of cell	P1	Mitochondria are known as power house of cell
P2	Energy required for various chemical activities	RP2	Energy necessary for various chemical activities
P3	Various chemical activities needed for life	RP3	Various chemical activities necessary for life
P4 hline	Energy released by Mitochondria	P4	Energy released by L1: P4 therefore P1
P5	Energy in the form of ATP molecules	P5	Energy in the form of ATP molecules
P6	ATP molecules known as energy currency of cell	P6	ATP molecules known as energy currency of cell
P7	Body uses Energy	P7	Body uses Energy
P5	Energy is stored in the form of ATP	P5	Energy is stored in the form of ATP
P8	ATP molecules used for making new chemical compounds	RP8	ATP molecules has role in synthesizing new chemical compounds
P9	ATP molecules used for performing mechanical work	RP9	ATP molecules has role in mechanical work
			L2: RP8 and RP9 therefore P6
P10	Mitochondria have two membrane coverings (instead of just one)	RP10	Mitochondria covered by two membrane coverings (instead of just one)
P11	Membrane consists of outer membrane	P11	Membrane consists of outer membrane
P12	Membrane consists of inner membrane	P12	Membrane consists of inner membrane
P13	Outer membrane is very porous	RP13	Outer membrane has form (very) porous
P14	Inner membrane is deeply folded	RP14	Inner membrane has form (deeply) folded
P15	Inner membrane creates large surface area	RP15	Inner membrane produces large surface area
			L3: RP14 therefore RP15
P16	Large surface area is used for ATP generating chemical reactions	RP16	Large surface area site for synthesis of ATP generating chemical reactions
P17	Mitochondria are strange organelles	RP17	Mitochondria known as strange organelles
P18	Mitochondria have their own DNA	RP18	Mitochondria consists of (their own) DNA
P19	Mitochondria have their own ribosomes	RP19	Mitochondria consists of (their own) ribosomes
			L4: RP18 and RP19 therefore RP17
P20	Mitochondria therefore can make their own proteins	RP20	Mitochondria can produce (their own) proteins
			L5: RP18 and RP19 therefore RP20

Table 3.2: List of re-represented propositions, with linking words emphasized.

Verbatim Sentences	Verbatim Propositions	Relations	Re-represented Propositions
S1: Each mitochondrion is a double membrane-bound structure with the outer membrane and the inner membrane.	P1: Each mitochondrion is a double membrane-bound structure	class-instance	RP1: Each mitochondrion is instance of double membrane-bound structure
	P2: Double membrane-bound structure with outer membrane and inner membrane	meronymy inclusion	RP2: Double layered membrane consists of outer membrane, inner membrane
S2: The leucoplasts are the colorless plastids of varied shapes and sizes with stored nutrients	P3: Leucoplasts are colourless plastids	class inclusion	RP3: Leucoplasts are (is-a) colourless plastids
	P4: Leucoplasts are of varied shapes and sizes	attribute	RP4: Leucoplasts have shape, have size varied
	P5: Leucoplasts are with stored nutrients	spatial inclusion	RP5: Leucoplasts contains stored nutrients

Table 3.3: Example showing the case of "is-a", "are" from content analysis of class 11 text.

Verbatim Sentences	Propositions	Relations	Re-represented Propositions
S1: Plastids also have their own DNA and ribosomes	P1: Plastids also have their own DNA and ribosomes	meronymy inclusion (component – integral object)	RP1: Plastids have part DNA and ribosomes Or its inverse RP1: DNA and ribosomes are part of Plastids
S2: The membrane of the erythrocyte has approximately 52 percent protein and 40 percent lipids	P2: Membrane of erythrocyte has 52% proteins and 40% lipids	meronymy inclusion (stuff – object)	RP2: Membrane of erythrocyte composed of 52% proteins and 40% lipids
S3: Mitochondria have two membrane coverings instead of just one	P3: Mitochondria have two membrane coverings (instead of just one)	meronymic inclusion (co-extensively)	RP3: Mitochondria covered by/surrounded by two membrane coverings (instead of just one)

Table 3.4: Example showing the case of "has/have" from content analysis from class 9 and 11 text. The linking word "has/have" is generally used for all kinds of inclusion relations – meronymic as well as spatial inclusion relations (Winston et al., 1987)

.

Verbatim Sentences	Propositions	Relations	Re-represented Propositions
S1: Peripheral proteins lie on the surface of membrane while the integral proteins are partially or totally buried in the membrane.	P1: Peripheral proteins lie on surface of membrane	spatial inclusion	RP1: Peripheral proteins located on surface of membrane
	P2: Integral proteins are partially or totally buried in the membrane	spatial inclusion	RP2: Integral proteins are partially or totally immersed in the membrane *proposed
S2: The leucoplasts are the colourless plastids of varied shapes and sizes with stored nutrients	P3: Leucoplasts are colourless plastids	class inclusion	P3: Leucoplasts are (is-a) colourless plastids
	P4: Leucoplasts are of varied shapes and sizes	property relation *proposed definition	RP3: Leucoplasts varies in shapes and sizes
	P5: Leucoplasts are with stored nutrients, Leucoplasts contains stored nutrients	spatial inclusion	RP4: Leucoplasts contains stored nutrients or Stored nutrients are contained in leucoplasts
S3: The nonpolar tail of saturated hydrocarbons is protected from the aqueous environment	P6: Non polar tail is protected from the aqueous environment	function *Nominalization	RP6: Non polar tail has role in /has function of protection from the aqueous environment

Table 3.5: Example showing the case of "of", "in", "on", "with", "about" from content analysis of class 11 text.

Domain, Attributes, Variable Properties	Range, Possible values
Color	Red, blue, violet,...
Height, Length, Width	10 meters, 5 meters,. . . .
Shape	Square, circular,. . .
PH	3, 7, 8, . . .
Temperature	Hot, cold,... 24C, 50C,. . .
Diameter	30nm, 50 microns, 5mm, 10cm, ...

Table 3.6: Examples depicting variable properties with possible values.

Concept term	linking word	concept term	concept term	attribute	concept term
Mitochondria	is	sausage-shaped or cylindrical shaped	Mitochondria	has shape	sausage shaped, cylindrical
Mitochondria	have	diameter of 0.2-1.0µm (average 0.5m)	Mitochondria	have diameter	0.2-1.0µm (average 0.5m)
Mitochondria	have	length 1.0-4.1µm	Mitochondria	have length	1.0-4.1µm

Table 3.7: Identifying and re-representing attributes from passage.

Verbatim Sentences (S)	Paraphrased Propositions (Pn)	Relations	Re-represented Propositions (RPn)
S1: The pili are elongated tubular structures made of a special protein.	P1: Pili are elongated tubular structures	Class inclusion	RP1: Pili IS-A structure
		Attribution	RP2: Pili have shape elongated, tubular
	P2: Pili are made of special proteins	meronymic inclusion	P2: Pili are made of special proteins
S2 The fimbriae are small bristle-like fibers.	P3: Fimbria are small bristle-like fibres	class inclusion	RP3: Fimbria IS-A bristle like fibre
		Attribution	RP4: Fimbria has length small

Table 3.8: Examples of Disambiguation of Attributes from content analysis of class 9 and 11 texts.

Re-represented propositions for attributes			RCM (precursor to process mo	
Concept	Linking word (attribute)	Concept (attribute value)	Attribute as Subject	Attribu
mitochondria	has diameter range	0.2 -1.0 micrometer	Diameter of mitochondria ranges from	0.2 -1.0
mitochondria	has length range	1.0-4.1 micrometer	Length of mitochondria ranges from	1.0-4.1
ribosomes	has form	dense	form of ribosomes is	denser
eukaryotic ribosomes	has unit	80S	unit of eukaryotic ribosomes is	80S

Table 3.9: The linking words are *nominalized* in subject place as a precursor for the representation of processes.

Structure Term	Linking Words	Structure Terms	Structure Terms	Process-centric lws	Process Terms
Mitochondria	releases	energy	Mitochondria	has role in	releasing energy
Leucoplasts	stores	materials	Leucoplasts	has role in	storage of materials
Vacuoles	provide	rigidity to cell	Vacoule	has role in	Provision of rigidity to cell
endomembrane system	regulates	protein traffic	endomembrane system	has role in	regulation of protein traffic
pancreatic cells	secretes	insulin	pancreatic cells	has role in	secretion of insulin
cell wall	protects	plant cells	cell wall	has role in	protection of plant cells
ribosomes	translates	mRNA into proteins	ribosomes	has role in	translation of mRNA into proteins
rough endoplasmic reticulum	makes	digestive enzymes	rough endoplasmic reticulum	has role in	synthesis of digestive enzymes

Table 3.10: Re-representing dynamic propositions from content analysis of class 9 and 11 texts.

Structure Term	Nominalized lws	Process Terms	Process Terms	Nominalized lws	Structure Terms
Mitochondria	has role in	releasing energy	Releasing of energy	is role of	Mitochondria
Leucoplasts	has role in	storage of materials	Storage of materials	is a function of	leucoplast
Vacoule	has a role in	Provision of rigidity to cell	Provision of rigidity to cell	is a function of	vacuole
Endomembrane system	has a role in	regulation of protein traffic	Regulation of protein	is the role of	endomembrane system
Pancreatic cells	has role in	secretion of insulin	Secretion of insulin	is function of	pancreatic cells
Cell wall	has a role in	protection of plant cells	Protection of plant cells	is the function of	cel wall
Ribosomes	has a role in	translation of mRNA into proteins	Translation of mRNA into proteins	is the role of	ribosomes
Rough endoplasmic reticulum	has a role in	synthesis of digestive enzymes	Synthesis of digestive enzymes	is a function of rough endoplasmic reticulum	

Table 3.11: Nominalization by re-arranging process as subject.

Process terms	Object terms	Variable Properties (Attributes)	Prior State (Values)	Post State (Values)
Folding of chromatin fibres	Chromatin Fibres	Appearance	loosely folded	densely folded
Folding of chromatin fibres	Chromatin Fibres	Length	longer	shorter

Table 3.12: Process representation as prior state and post-state of content from PG text.

Processes	Objects	Variable Properties (Attributes)	Values (Prior State)	Values (Post State)
Disintegration of nucleoli	Nucleoli	Appearance	Prominent	Less Prominent, Fragmented
• the appearance of nucleoli becomes fragmented				
Reduction in Nucleoli	Nucleoli	Size	Larger	Smaller
• the size of nucleoli becomes smaller				
Fragmentation of nuclear membrane	Nuclear Membrane	Appearance	Intact	Fragmented
• the appearance of nuclear membrane becomes fragmented				
Lengthening of Spindle Fibres	Spindle Fibres	Size	Shorter	Larger
• the size of spindle fibers becomes longer				
Changes in Cell	Cell	Shape	Round	Spheroid
• the shape of cell becomes spheroid				

Table 3.13: Process representation of prophase depicting change as prior state and post state from content of PG text.

Object/Process Terms	Linking Words	Examples
Describing a role or a function	has role in, has function	ribosome has role in protein synthesis
A sub-process	has sub-process	photosynthesis has subprocesses light reactions, calvin cycle
A simultaneous process	occurs simultaneously with	chromatin condensation occurs simultaneously with nuclear membrane fragmentation
Sequences or cyclical processes	followed by	glycolysis is followed by fermentation
Cause-effect processes	causes, has effect, results in	vitamin B12 deficiency results in anemia
changes to, Developmental processes	becomes, developed into transformed into	egg developed into larva

Table 3.14: Examples depicting linkages for processes.

Code Number	Category	Examples
	Relations	
C1	Meronymic inclusion	has part, composed of
C2	Class inclusion	is a
C3	Class member	instance of
C4	Spatial inclusion	covered by, located in
C5	Function	has role, has function
C6	Process	changes into, develops, becomes, results into
	Attributes	
C7	Geometric Form	shape, appearance
C8	Measurable	size, length, diameter, width
C9	Quality	color

Table 3.15: Compiled List of Semantic Categories of Relations and Attributes.

Rule Number	Re-representation Rules
R1	Pass those statements that already use the object and data properties available in the Reference Set appropriately
R2	Eliminate ambiguity of common verbs, is, are, restricting their lone usage to an unambiguous category of semantic relation for subtyping or use them only in combination with the object and data properties available in the Reference Set
R3	Eliminate ambiguity of common verbs, has, and have, restricting their lone usage. Based on the intended meaning, use verbs has, and have always in combination with data property, or object property as appropriate
R4	Eliminate the lone usage of prepositions as semantic relations between subject and object, and use them only in combination with the object and data properties available in the Reference Set and to create complex expressions between class terms
R5	Rewrite the adjectives in the form of data properties or its inverse names such as size, color, shape, length, form, density, pH, viscosity, solubility, etc.
R6	Rewrite the verbs that represent processes/events into nouns, by using object properties such as has role/has function, has sub-process, is site of, etc.
R7	Rewrite the verbs in the form of object properties or their inverse names, after nominalization when necessary
R8	Represent the change in the events as prior-state and post-state following temporal order
R9	Add prepositions to create condensed expressions
R10	Relating statements through logical relations such as because, therefore, and, hence, or, etc.

Table 3.16: Compiled List of Re-representation Rules.

Chapter 4

Semantic Content Analysis Using Re-representation: A Methodology

4.1 Introduction

Content analysis broadly involves coding, categorizing, comparing, and concluding (Cohen et al., 2015, p. 564). It incorporates several steps as outlined by Cohen et al. (2017): defining research questions; defining the population from which units of text are to be sampled; defining the sample; defining the context; defining the units of analysis; deciding the codes for analysis; construct the categories for analysis; conduct the coding and categorizing of data; conduct content analysis; summarizing findings; making speculative inferences. We have adopted these steps for the content analysis of the science texts as they are suitable for our study.

4.2 Thesis Objective - 2

To demonstrate the use of the Re-representation method for semantic content analysis of biology text, focusing on the linking words. For this, we apply the content analysis methodology for the analysis of 6 textbooks as shown in Figure 4.1.

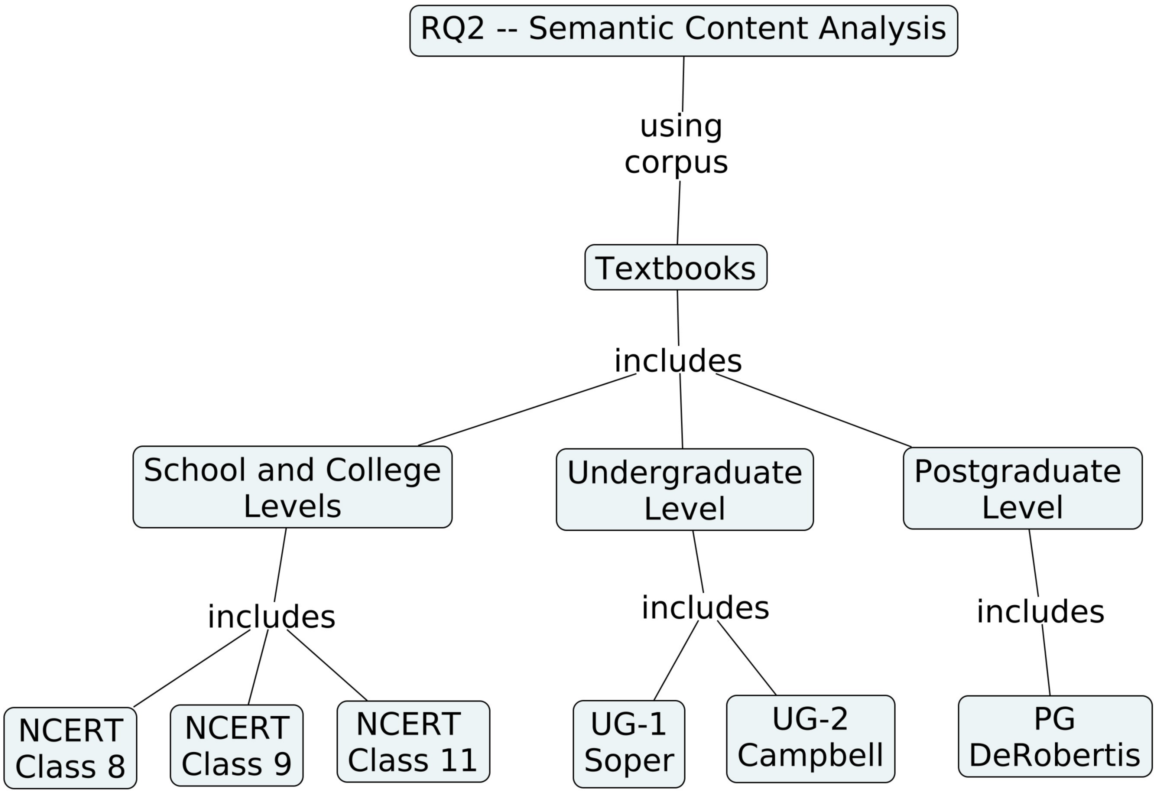


Figure 4.1: A graphical representation of the research model for research question 2. (To read from top to bottom and left to right).

4.3 Textbook as Corpus

Content analysis can be worked on any written material such as text, documents, interviews, transcripts, audios, videos, etc. The domain of analysis of our study is the text. Textbooks are an integral part of any education teaching-learning system and are a primary aid to help students obtain knowledge. These are used not only as a pedagogical tool but also as a research tool in science education. Several researchers have used textbook analysis for specific purposes – studying the nature of science (Chiappetta et al., 2004; McComas, 2014; Niaz and Maza, 2011); analyzing texts on history and philosophy of science (Niaz and Maza, 2011); analyzing diagrammatic representations (Thiele and Treagust, 1995), etc. Textbooks represent school discipline to students, therefore, it is inevitable that adequate and consistent scientific knowledge is presented in the science textbooks, or else it can negatively affect students' ideas. Textbooks as source were also part of research studies for comparison of automatic extraction of concept maps with human-generated concept maps (Olney et al., 2011). In the recent past, textbooks have been also used for knowledge representation in education (Chaudhri et al., 2014b) specifically focusing on life science textbooks which is also parallel with our subject of study.

Educationists have pointed out some criticism of textbooks concerning reading difficulty, comprising of difficult vocabulary, uninteresting language, lack of organization and style, outdated and filled with errors, jargonified scientific terms, and inaccuracy of information in textbooks. Considering the criticism of the use of uninteresting language in textbooks, it is mentioned that textbooks are being shortened to fit into a readability formula, thus compromising the explicit and vivid writing style. During the analysis of textbooks for language Muspratt and Freebody (2013) observed systematic variation when deploying linguistic features while representing scientific knowledge. At times, the syntactic and semantic features are dissolved, thus resulting in incoherent, more abstract language.

Despite the criticisms, textbooks play a central role as these are the links between scientific knowledge and students' understanding. As our focus is on seeking patterns and analyzing text, we use the textbook as the corpus of our study.

4.3.1 Choice of Form

The biological knowledge is usually described in propositional statements for explanation. In addition to this, symbolic representations (equations, graphs), diagrams, pathways, cycles, photographs, maps are often used in textbooks for expression and justification of biological knowledge. Often at school or college level textbooks, use of historical accounts, metaphors and analogies are also part of expression of a phenomena. Use of models, and model-based reasoning as representational tools has a crucial role in representation of biological knowledge. From these diverse aspects of biological knowledge, we have limited to only propositional statements that describe structure and processes.

The other forms of biological knowledge such as equations, graphs are unambiguous and are rigorous and therefore it was not selected as part of the study. The metaphors and analogies were also not considered for the study.

4.3.2 Sampling of Textbooks

The sampling of texts is to be done based on key issues of representativeness, access, size of the sample, and generalizability of the results (Cohen et al., 2015). The sampling strategy followed for texts is similar to those followed in a sampling of the population. Purposive sampling is often considered as a key feature of qualitative research. We used a purposive sampling to identify the textbooks to be analyzed, targeting the cell biology domain of school and college levels as shown in Table 4.1.

Sr. No.	Class/Levels	Textbooks	Chapter/Topics	Pages
		School and High School Levels		
1.	8	Science (NCERT, 2007b)	Chapter 8: Cell: Structure and Functions	92-101
2.	9	Science (NCERT, 2007c)	Chapter 5: The Fundamental Unit of Life	57-67
3.	11	Biology (NCERT, 2007a)	Chapter 8: Cell: The Unit of Life	125-141
		Undergraduate Levels		
4.	Undergraduate (UG-1)	Biological Science (Taylor et al., 2003)	Chapter 5: Cells	128-143
5.	Undergraduate (UG-2)	Biology (Campbell and Reece, 2008)	Chapter 6: A Tour of the Cell	98-120
		Post graduate Levels		
6.	Postgraduate (PG)	Cell and Molecular Biology (De Robertis and De Robertis, 1985)	Chapter 15: Mitosis and Cell Division	420-423

Table 4.1: Details of Text Corpus.

The school textbooks were chosen based on their wider usage, ease of availability across the country, and being developed by a central agency, the National Council of Educational Research and Training (NCERT) (NCERT, 1961). The NCERT is an apex resource organization that assists and advises the Central and State Governments on academic matters related to school education. The NCERT provides academic and technical support for qualitative improvement of school education and undertakes research, development, training, publication programs, etc. (NCERT, 1961). One of its prime objectives is that the NCERT develops and publishes textbooks from Class I upto Class XII in all subjects. The current textbooks are developed based on the National Curriculum Framework (NCERT, 2005). Another reason for the selection of sample textbooks is that these are used all over the country where science is taught in English as a medium of instruction and examinations are conducted by the Central Board of Secondary Education (CBSE) (CBSE, 1962) in India.

The sampling of textbooks (see Table 4.1) was done based on curriculum, grades, chapter, and content. We wanted to focus on the middle, secondary, and higher secondary levels, and hence, school textbooks from Classes VIII (NCERT, 2007b), IX (NCERT, 2007c) and XI (NCERT, 2007a), catering to the student age of 13-17, were chosen for the study. Further, to extend the focus on the college level, three textbooks were also chosen from pre-university (Taylor et al., 2003), undergraduate (Campbell and Reece, 2008), and postgraduate (De Robertis and De Robertis, 1985) levels.

We now discuss how the textbooks at each level describe the topics. In classes VIII and IX, all the sciences (physics, chemistry, and biology) are presented in one textbook each. Class XI has separate textbooks, each for physics, chemistry, and biology. Within these three textbooks of school levels, we selected the chapters: "Cell: Structure and Function" for Class VIII; "The Fundamental Unit of Life" for Class IX; and "Cell: The Unit of Life" for Class XI. The college textbooks considered for the study are used in many countries in college biology education. The chapters considered are: "Cells" (Taylor et al., 2003); "A Tour of the Cell" (Campbell and Reece, 2008); and "Mitosis and Cell Division" De Robertis and De Robertis (1985). The corpus of 6 textbooks is shown in Table 4.1.

4.3.3 Data, Topics, Content

As our interest is in biology texts, we selected a fundamental topic that spans from high school and college to undergraduate levels. The content for each chapter on cell biology was broken down into several sub-topics, such as the nucleus, mitochondria, cytoplasm ribosomes, etc. Each topic was analyzed for sentences and the propositions as shown in Table 4.2.

Textbooks	Topics	Sentences	Verbatim Propositions	Re-represented Propositions
8	10	151	266	262
9	13	131	258	330
11	22	225	498	476
UG-1	18	329	602	618
UG-2	20	306	541	698
PG	7	16	45	75
Total Corpus	90	1158	2205	2458

Table 4.2: Total corpus data from textbooks.

4.4 Qualitative Study

We employed a qualitative research methodology to gain insights into linking words in representing knowledge in the text. The characteristics of qualitative research that it: "explores a problem and develops a detailed understanding of a central phenomenon" (Creswell, 2012, p. 16) have enabled us to adopt it. Qualitative research is a form of social inquiry focusing on interpretations and making sense of experiences. Qualitative research is defined as "multi-method in focus, involving an interpretive, naturalistic approach to its subject matter" (Denzin and Lincoln, 1998). Qualitative researchers study things in their natural settings to interpret phenomena in terms of their meanings.

Qualitative research can be done in several forms: content analysis, grounded theory, phenomenology, hermeneutics, ethnography, interpretive, etc. To address the research study, we have chosen 'content analysis' of the text as a specific methodology. The most significant feature of content analysis is to seek patterns and trends in textual data (Weber, 1985). Content analysis is an unobtrusive technique (Krippendorff, 2000), that does not require the researcher to intrude in the research context. Content analysis focuses on language and meaning in context, and it is systematic and verifiable since the code categories are explicit (Mayring, 2014). Through content analysis, verification and replication are possible, as the data are in a permanent form (texts) that can be reanalyzed if required.

The qualitative content analysis methodology was adopted to gain insights into knowledge representation. The aim was not to measure or quantify textbook knowledge but to understand the textbook's representation of knowledge. The researcher's role was to conduct the content analysis to categorize linking words to gain insights and seek patterns into how the text represents knowledge.

The research questions of our study drive the methodology towards qualitative content analysis. The objective of the thesis is focused on extracting and interpreting the meaning of the choice of linking words in science texts. These objectives are contrasted with quantitative methods wherein the primary intent is testing hypotheses and establishing relations between variables. It was the researcher's standpoint that quantitative methods were unlikely (unsuitable) to elicit the meaning from the rich data to address the central phenomena of the study. In the researcher's view, the study's underlying assumptions enabled a qualitative stance to fit with this study. Employing qualitative content analysis methodology in research makes it more exciting to enjoy the serendipity and to discover patterns. As our objective is to seek patterns and identify trends as an outcome of analyzing the text, content analysis was considered the most suitable methodology.

4.4.1 Content Analysis

The content analysis comprises "a strict and systematic set of procedures for the rigorous analysis, examination, and verification of the contents of written data" (Cohen et al., 2015, p. 563). Content analysis is a method of observation in which, instead of asking people to respond to questions, it "takes the communications that people have produced and asks questions of communications" (Kerlinger, 1986). Therefore, it is also considered an unobtrusive or non-reactive method of analysis. Krippendorff (2004) defines content analysis as a research technique that utilizes procedures for making replicable and valid inferences from data to their context. Weber (1985) states that it is a process by which ‘many words of texts are classified into much fewer categories’ (Weber, 1985). The content analysis is about making valid, replicable, and objective inferences about the text based on explicit rules.

Through content analysis, the researchers can analyze the text and summarize it into both pre-existing categories (inductively) or emerging categories (deductively) to generate or test a theory. Content analysis involves coding, categorizing, comparing, and drawing conclusions from the text. Additionally, some other essential features of content analysis are examining the interconnectedness of categories, seeking the emergent nature of themes or patterns, and developing a theory.

Content analysis aims to convert raw text into data that can be analyzed scientifically to form a body of knowledge. Content analysis by itself conforms to at least two fundamental principles of the scientific method – objectivity and generalizability. Objectivity comes by adopting explicit rules during analysis, enabling different researchers to obtain the same results from the same text. Generalizability can be achieved when results obtained in a given domain can also be applied to other domains. Content analysis offers several advantages, such that it refers directly to communication via texts and hence gets at a central phenomenon of analysis. It can be used to develop expert systems, as the coding involves explicit rules for analyzing relations between concepts. Although a qualitative research method, content analysis can help to make a quantitative expression of a pattern, trend, or phenomenon.

In our current study, content analysis is used as a research method to determine the choice and use of linking words within texts from topics in chapters. During content analysis, the text is coded into categories on various levels—passages, sentences, propositions, and linking words. These are then examined using relational analysis, one of the primary forms of content analysis.

4.5 Relational Content Analysis

Relational analysis, although like conceptual analysis, starts with identifying concepts in the unit of analysis. However, it goes beyond concepts by exploring the relations or links between the concepts. Relational analysis is also called semantic analysis (Palmquist and Dale, 1997). In other words, relational analysis focuses on the semantic links between concepts. The rationale for choosing relational analysis comes from the theoretical perspective of semantic holism. Concepts have no inherent meaning; meaning is a product of the relationships between concepts in a text. Concepts acquire meaning by connecting to other concepts (Carley, 1990). A known advantage of relational analysis is that it can help achieve a higher degree of computational analysis. It is claimed that perhaps relational analysis can maintain a higher degree of statistical rigor without losing the richness of details apparent in qualitative methods.

The methodology used for the study is relational content analysis of scientific texts (Cohen et al., 2015; Carley, 1990; Creswell, 2008). The corpus of content analysis is six texts of increasing subject complexity on cell structure and function at school, college, and undergraduate levels. We adopted the textbook format for content analysis because, as a written medium, it is widely used in schools and colleges. Well-chosen authentic subject texts seem intuitively to be a sound and practical way to access and analyze. The textbooks are also more feasible for obtaining comprehensive and large-scale data. The purposive sampling method was employed to choose biology textbooks (for cell biology topics).

The relational content analysis was performed on the sentences. As the content analysis focuses on the choice of linking words used, the sentences were first paraphrased into propositions. Then, the linking words were coded, categorized and analyzed for patterns. All the codings, categorizations, analyses, and propositions have been manually generated in our study.

4.5.1 Units of Analysis

The units of analysis can be at many levels ranging from word, phrase, sentence, paragraph, text, etc. (Cohen et al., 2017). In the context of our study, the sentences are the units of analysis. According to Krippendorff (2000), the unit of analysis can be classified into three kinds – sampling units, recording/coding units, and context units. The sentences, considered syntactical, are the sampling units. The sentences are first paraphrased to obtain a node-link-node or concept-linking word-concept format, wherein the nodes are the concepts (class terms), and links are the relations and attributes (object properties and data properties) that connect the concepts. Usually, the concepts belong to scientific terms (nouns), whereas the linking words belong to natural language vocabulary (verb/adjective).

4.5.2 Coding Units

Due to the focus of our study on relations, the coding units are the 'linking words.' The sentences are paraphrased into propositions in the concept-linking word-concept format. The proposition with verbatim linking words (LW) is the verbatim proposition, while the proposition with re-represented linking words (RLW) is the re-represented proposition. Table 4.3 illustrates the linking words with emphasis.

4.5.3 Categories

Categorization of the content is crucial in the process of content analysis. It is fundamental in learning, decision-making, etc. The coding units i.e., "linking words" are categorized for some standard features.

Classically, Aristotelian criteria for categorization comprises a set of properties with sufficient conditions to capture meaning and be shared by its members. The classical research in cognitive psychology (Smith and Medin, 1981) has applied the categorization of objects into superordinate, basic-level, subordinate categories (Markman and Wisniewski, 1997), as prototypes (Mervis and Rosch, 1981), etc. A basic level category, such as

Sentence	Concepts	Linking Words	Concepts
Algae have cell wall, made of cellulose, minerals	Algae	*have*	cell wall
	Algae	*covered by*	cell wall
	Cell wall	*made of*	cellulose
	Cell wall	*made of*	minerals
Ribosomes are about 15 nm by 20 nm in size	Ribosomes	*are about*	15 nm by 20 nm in size
	Ribosomes	*have size*	15 nm by 20 nm

Table 4.3: Coding units are emphasized for relations and attributes for both verbatim and re-represented propositions.

a book, can have a general category (superordinate), such as reading material, and a specific category (subordinate), such as fiction. As per Ausubel et al. (1978) classroom learning theory of subsumption, new concepts are learned by subsuming with the existing concept based on prior learning. For example, 'bird' is a current concept, and 'tweety' a new concept, is learned by subsumption of 'tweety is a bird'. These studies indicate the significance of categories in learning and representing knowledge. We also follow the cascading or hierarchical basis of categories.

For our study, we established the following categories wherein each coding unit, i.e., linking words, was grouped into the appropriate category depending on its intended meaning. The categorization of relations (LW or RLW) manifests at three levels, as shown in Figure 4.2 for relations and in Figure 4.3 for attributes.

The first and foremost is to categorize the coding unit of linking words into relation or attribute assigned as level 1. Further, the linking words of level 1– relation – will be further categorized into – meronymy-inclusion – as level 2 and into – part-whole or material-composition – as level 3. Similarly, the linking words of level 1 – attribute – will

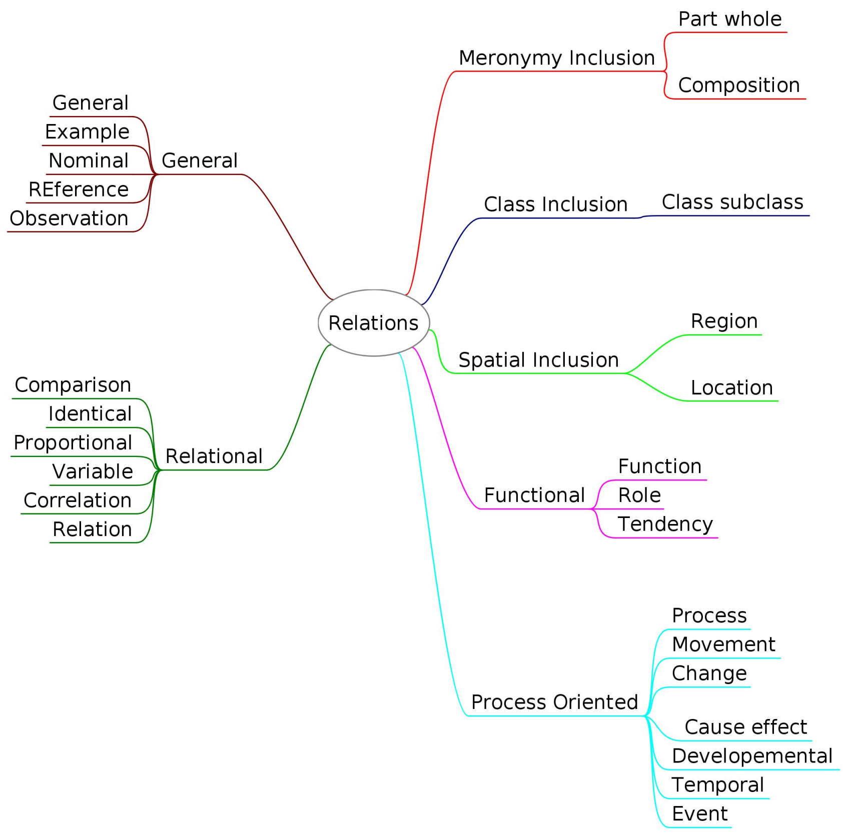


Figure 4.2: Categories for Relations.

be further categorized into – measurable – as level 2 and further into – size or length – as level 3. We show the categorizing of coding units (linking words) into level 1, level 2, and level 3 in Table 4.4.

4.6 Procedure for Semantic Content Analysis

It is to be noted that the content analysis method is applied to the propositions for all of the data for the thesis work. We also emphasize that during the analysis processes, the concept terms are not deleted nor newly added. Sometimes, the concept terms may be changed from a structure term to a process term due to nominalization (Halliday, 2006). As far as possible, the body of knowledge is preserved. This is discussed in chapter 5, Figure 5.12.

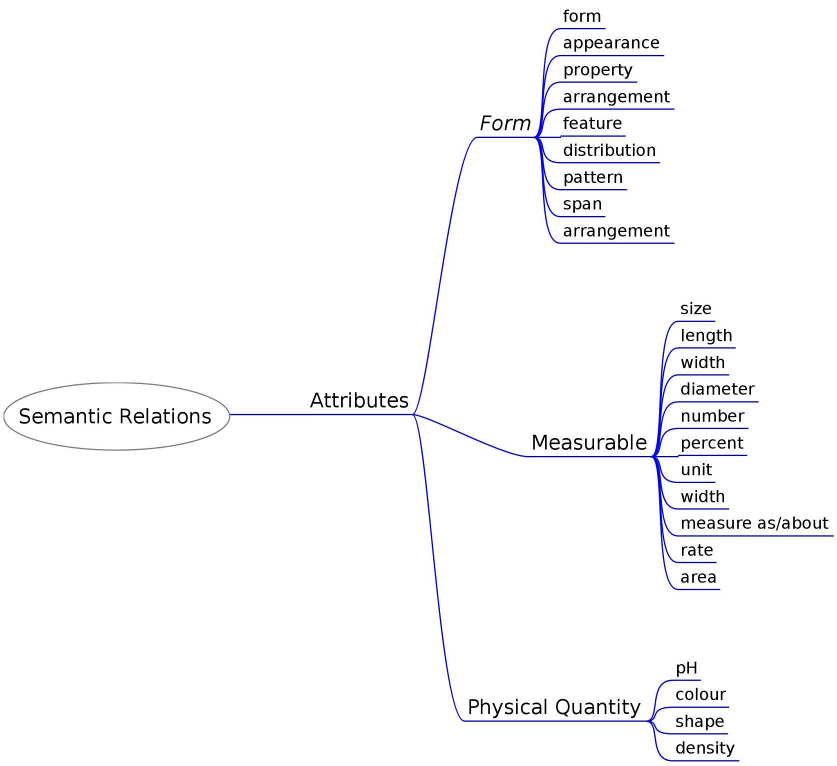


Figure 4.3: Categories for Attributes.

We describe and illustrate in detail the procedure of semantic content analysis followed for both verbatim and re-represented propositions. As our early explorations of the research began with the concept mapping method, we represented the text using the Standard Concept Mapping (SCM) method and then re-represent it with the Re-represented Concept Mapping (RCM) method.

The detailed steps followed in the procedure for semantic content analysis are shown below:

4.6.1 Paraphrasing

The text passages are comprised of sentences with concepts and linking words. The passages, often containing simple or compound sentences, are broken down into single sentences with the syntax of subject-predicate-object as shown in Figure 4.4. This is in line with the Re-representation Phase-1 method explained in Chapter 3. Even in the

Coding Units			Categories		
Concepts	Linking Words	Concepts	Level 1	Level 2	Level 3
The vacuole is a membrane-bound space found in the cytoplasm					
Vacoule	is a	membrane-bound space	Relation	Class inclusion	Class-subclass
Vacoule	found in	Cytoplasm	Relation	Spatial inclusion	Region
Mitochondria is sausage-shaped or cylindrical, having a diameter of 0.2-1.0 micrometer and length 1.					
Mitochondria	has shape	Sausage-shaped or cylindrical	Attribute	Physical Quantity	Shape
Mitochondria	has diameter	0.2-1.0 micrometer			
Attribute Measurable	Size				
Mitochondria	has length	1.0-4.1 micrometer	Attribute	Measurable	Length

Table 4.4: Categorization of relation and attribute into level 1, level 2 and level 3.

studies related to creating Medical Wordnet from WordNet database, paraphrasing is used for translating online content into basic propositions (Smith and Fellbaum, 2004). Table 4.5 illustrates a few examples of paraphrasing textbook sentences into propositions (columns 1 and 2). Chapter 3, discusses about paraphrasing for converting sentences into propositions.

4.6.2 Re-representing

The re-representation of scientific text is a unique part of the research work (chapter 3) and, therefore, in the semantic content analysis. The re-representation is of linking words alone, only when necessary, and not of the concepts. Whenever there is an instance of a vague, ambiguous linking word or mere use of a preposition then we replace the

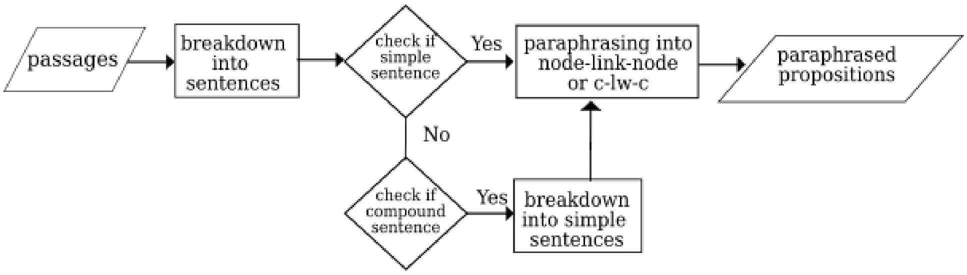


Figure 4.4: Paraphrasing Sentences into Propositions.

linking word with well-defined linking words from the Reference Set (see Chapter 3). The Reference Set is used to validate re-represented linking words and attributes. The method and decision applied to re-represent propositions are shown in Figure 4.5.

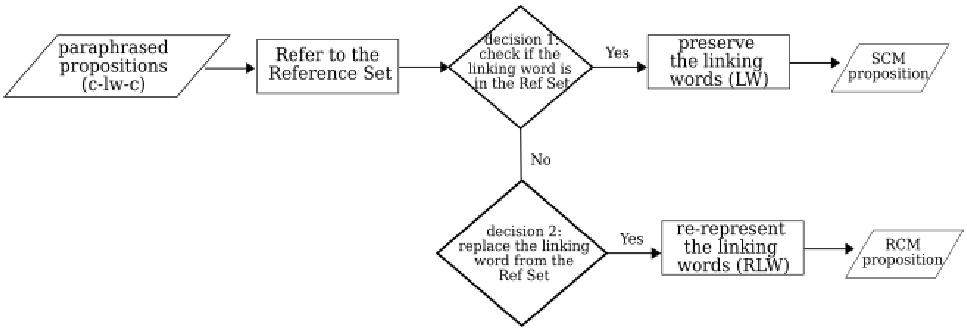


Figure 4.5: Re-representing Propositions.

The decision whether to replace or maintain the linking words is considered based on the intended meaning of the proposition, usage of linking words. The Table 4.5 illustrates the re-representation of linking words from paraphrased propositions to re-represented propositions (columns 2 and 3) only when required.

4.6.3 Assigning Coding Units

The coding units are assigned to the paraphrased propositions of subject-predicate-object, or concept-linkingword-concept as shown in Figure 4.6.

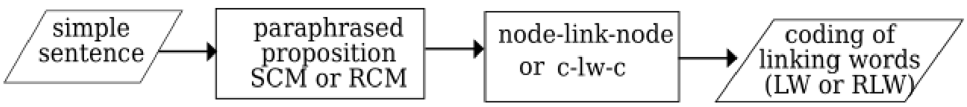


Figure 4.6: Assigning Coding Units.

Table 4.6 illustrates the process of assigning of coding units to both the paraphrased and re-represented propositions.

4.6.4 Categorizing the Coding Units

Categorization of the content is crucial in the process. An exhaustive list of categories is required to address content validity. Constructing categories involves grouping units into groups of semantic relations. For the study, we established three steps of categorization wherein each coding unit was grouped into the appropriate category depending on its implicit and intended meaning.

4.6.5 Categorizing for Relations and Attributes – Level 1

The initial categorization of linking words of level 1 is based on the decision of whether these belong to a relation or attribute, as shown in Figure 4.7. These are mutually exclusive categories, as shown in Table 4.7. An elaborate discussion and distinction between relation and attribution is discussed in Chapter 3.

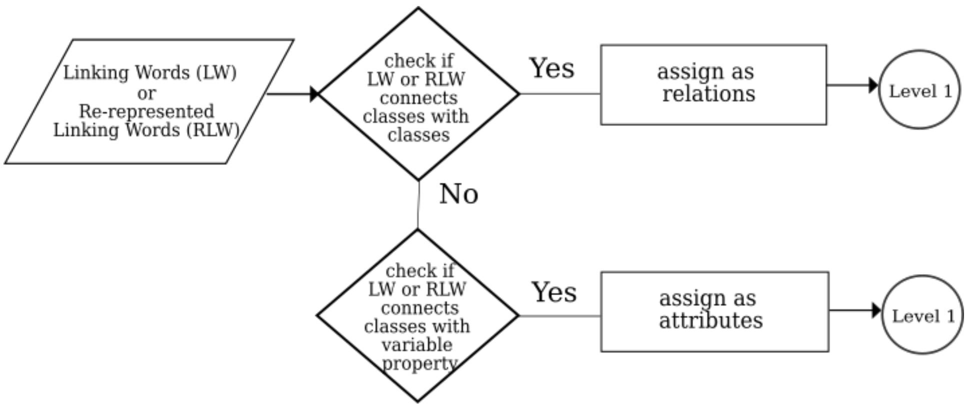


Figure 4.7: Categorization – Level 1.

4.6.6 Categorizing for Dimension – Level 2

When the linking words belong to the relations, these are further categorized as Level 2 based on the role and dimension as meronymy inclusion, class inclusion, spatial inclusion, functional, process-oriented, relational, general (Winston et al., 1987; Smith et al., 2005) as shown in Figure 4.8. The attributes are also categorized as Level 2: form, measurable property, and physical quantity.

The Table 4.8 illustrates the categorization of linking words into level 2.

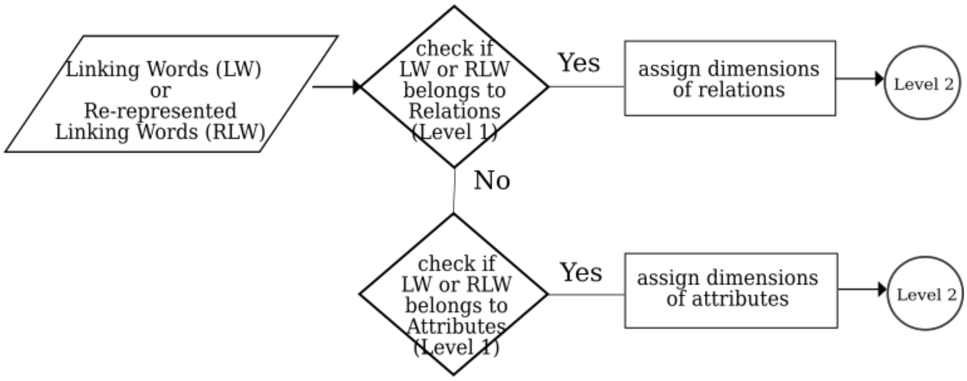


Figure 4.8: Categorization – Level 2.

4.6.7 Categorizing for Specific Kind – Level 3

The linking words after each of Level 2 categories are further categorized into specific Level 3 categories as shown in Figure 4.9 for relations and attributes. The meronymy relations include part-whole and material-composition; the spatial inclusion relation includes location and region.

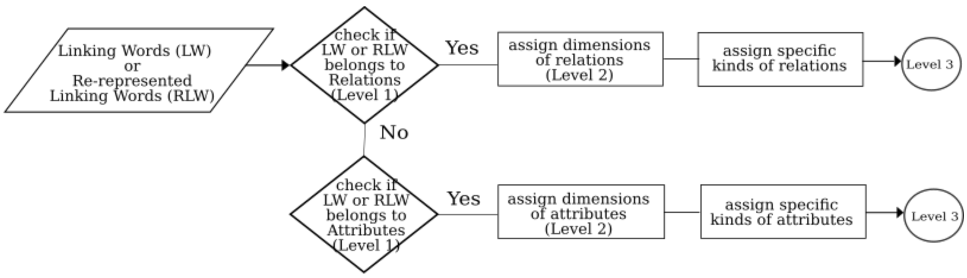


Figure 4.9: Categorization – Level 3.

Table 4.9 illustrates the Level 3 categorization of relations. However, these two linking words belonging to meronymic inclusion have different meanings: “has part” is related to a structure-structure relation, while “made of” is associated with a structure-composition relation. These finer distinctions are taken care of while categorizing for Level 3.

With the assigning of categories in Levels 1, 2, 3, the content analysis procedure is completed, atleast for the work to be done in this study. Here we would like to mention that these categories for the domain of the study seems sufficient. These categories need not be exhaustive and can be extended depending on another domain of study.

4.7 Summary

In this chapter, we have provided a methodology for using the developed Re-representation method for semantic content analysis of scientific text. Relational content analysis was used to illustrate the categorization of linking words for representing knowledge. The sampling of texts was of 6 purposefully selected textbooks with increasing levels of complexity. The data was coded and categorized based on the three levels.

Further, once we characterize the linking words, we seek patterns to make certain inferences from the semantic content analysis (research question 2). We discuss the research question 2 in the thesis using the re-representation method and its data analysis and findings in chapter 5.

Verbatim Sentences	Paraphrased Propositions	Re-represented Propositions
The Golgi apparatus, first described by Camillo Golgi, consists of a system of membrane-bound vesicles	Golgi apparatus described by Camillo Golgi	Golgi apparatus described by Camillo Golgi
	Golgi apparatus consists of membrane-bound vesicles	Golgi apparatus consists of membrane-bound vesicles
	Membrane-bound vesicles are part of endomembrane system	Membrane-bound vesicles are part of endomembrane system
Ribosomes are associated with the plasma membrane of the cell. They are about 15 nm by 20 nm and are made of two subunits - 50S and 30S units of prokaryotic ribosomes. Ribosomes are the site of protein synthesis.	Ribosomes associated with plasma membrane	
	Ribosomes are about 15nm by 20 nm in size	Ribosomes have size about 15nm by 20 nm
	Ribosomes made of two subunits	Ribosomes made of two subunits
	Two subunits are 50S, 30S units	Two subunits includes 50S, 30S units
	Ribosomes are site of protein synthesis	Ribosomes are site of protein synthesis
Plastids also have their own DNA and ribosomes	Plastids have DNA	Plastids consists of DNA
	Plastids have ribosomes	Plastids consists of ribosomes
The interphase nucleus has nucleoprotein fibres called chromatin, nuclear matrix and spherical bodies called nucleoli.	Interphase nucleus has nucleoprotein fibres	Interphase nucleus consists of nucleoprotein fibres
Nucleoprotein fibres are called chromatin	Nucleoprotein fibres are called chromatin	
	Interphase nucleus has nuclear matrix	Interphase nucleus consists of nuclear matrix
	Interphase nucleus has spherical bodies	Interphase nucleus consists of spherical bodies
	Spherical bodies are called nucleoli	Spherical bodies are called nucleoli

Table 4.5: Sentences paraphrased into propositions and re-represented propositions.

Re-represented Propositions	Concepts	Linking Words (Coding Units)	Concepts
Golgi apparatus consists of membrane-bound vesicles	Golgi apparatus	consists of	membrane-bound vesicles
Membrane-bound vesicles are part of endomembrane system	Membrane-bound vesicles	part of	endomembrane system
Ribosomes associated with plasma membrane	Ribosomes	associated with	plasma membrane
Ribosomes have size about 15nm by 20 nm	Ribosomes	have size	about 15nm by 20 nm
Ribosomes made of two subunits	Ribosomes	made of	two subunits
Two subunits includes 50S, 30S units	Two subunits	includes	50S, 30S units
Ribosomes are site of protein synthesis	Ribosomes	are site of	protein synthesis
Plastids consists of their own DNA	Plastids	consists of	their own DNA
Plastids consists of their own ribosomes	Plastids consists of	their own ribosomes	
Interphase nucleus consists of nucleoprotein fibres	Interphase nucleus	consists of	nucleoprotein fibres
Nucleoprotein fibres are called chromatin	Nucleoprotein fibres	are called	chromatin
Interphase nucleus consists of nuclear matrix	Interphase nucleus	consists of	nuclear matrix
Interphase nucleus consists of spherical bodies	Interphase nucleus	consists of	spherical bodies
Spherical bodies are called nucleoli	Spherical bodies	are called	nucleoli

Table 4.6: Assigning Coding Units.

Concepts	Linking Words (Coding Units)	Concepts	Categorization- Level 1
Golgi apparatus	consists of	membrane-bound vesicles	Relation
Membrane-bound vesicles	part of	endomembrane system	Relation
Ribosomes	associated with	plasma membrane	Relation
Ribosomes	have size	about 15nm by 20 nm	Attribute
Ribosomes	made of	two subunits	Relation
Two subunits	includes	50S, 30S units	Relation
Ribosomes	are site of	protein synthesis	Relation
Plastids	consists of	their own DNA	Relation
Plastids consists of	their own ribosomes	Relation	
Interphase nucleus	consists of	nucleoprotein fibres	Relation
Nucleoprotein fibres	are called	chromatin	Relation
Interphase nucleus	consists of	nuclear matrix	Relation
Interphase nucleus	consists of	spherical bodies	Relation
Spherical bodies	are called	nucleoli	Relation

Table 4.7: Categorizing Coding Units – Level 1.

Linking Words (Coding Units)	Categorization- Level 1	Categorization - Level 2
Consists of	Relation	Meronymic inclusion
Part of	Relation	Meronymic inclusion
Made of	Relation	Meronymic inclusion
Includes	Relation	Class inclusion
Have size	Attribute	Measurable
Have shape	Attribute	Physical Quantity
Have length	Attribute	Measurable
Have colour	Attribute	Physical Quantity

Table 4.8: Categorizing Coding Units – Level 2.

Linking Words (Coding Units)	Categorization- Level 1	Categorization - Level 2	Categorization - Level 3
Consists of	Relation	Meronymic inclusion	Part-whole
Part of	Relation	Meronymic inclusion	Part-whole
Made of	Relation	Meronymic inclusion	Composition
Includes	Relation	Class inclusion	Class-subclass
Have size	Attribute	Measurable	Size
Have shape	Attribute	Physical Quantity	Shape
Have length	Attribute	Measurable	Length
Have colour	Attribute	Physical Quantity	Colour

Table 4.9: Categorizing Coding Units – Level 3.

Chapter 5

Semantic Content Analysis: Findings, Outcomes and Patterns

5.1 Introduction

Developing the re-representation method has been one of the objectives of the thesis. The details of the re-representation method are presented in Chapter 3. It is of significance to analyze what patterns can emerge as a result of the re-representation method. In this chapter, we present the findings. It is to be noted that the Phases 1-4 have been applied for the semantic content analysis and findings.

The main essence of semantic content analysis using the developed re-represented method has been seeking patterns, categories, themes, and regularities from school, college, and undergraduate textbook data to address research question 2. The focus has been on the nature and kinds of linking words (LW) and the re-representation of linking words (RLW).

The methodology illustrated in Chapter 4 is followed throughout the data analysis. We begin with the corpus of the data and demonstrate a sample analysis.

5.2 Characterization and Re-representation

As our work focuses on linking words, we first illustrate the characterization method and the re-representation of linking words, for relations and attributes.

Table 5.1 shows the categorization of relations into dimensions.

Dimensions of relations	Specific categories of Relations	Examples
Meronymy	Part-whole	consists of, composed of
Class inclusion	class subclass	includes
Spatial inclusion	location, region	surrounded by, contains
Function	Function, role, tendency	has role, has function
Process	Change, Development, movement, process, temporal	keeps changing, developed by, moves across, undergoes process, during
Causal	Cause-effect	caused by, results into
Others	Comparison, conditional, identical, nominal, observation, reference, relational, variability	Greater than, needed for, same as, called as, observed, represents, related to, varies in

Table 5.1: Dimension-wise categorization of linking words (relations).

Table 5.2 shows the categorization of attributes into dimensions.

Dimensions of Attributes	Specific categories of Attributes	Examples
Measurable	Size, Length	has size, has length
Form	Appearance	has property
Physical Quantity	Shape, Colour	has shape, has colour

Table 5.2: Dimension-wise categorization of linking words (attributes).

5.3 Sample Data Analysis

We follow the steps for the data analysis as shown in Figure 5.1.

We categorize the linking words, aggregate and collate the linking words dimension-wise, and count the kinds of linking words and the frequency of each linking word. This is

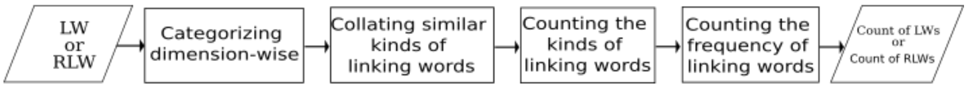


Figure 5.1: Series of steps followed in data analysis.

followed for each topic, and then a cumulative analysis is done for each level of text on the paraphrased propositions (P) and the re-represented propositions (RP).

We illustrate the steps of analysis using data in Tables 5.3, 5.4, 5.5, 5.6, 5.7.

- We first begin with counting the verbatim linking words (LW) of the entire data of all 5 textbooks. A sample of meronymic inclusion dimension is shown in Table 5.3.
- Similarly, for re-represented linking words (RLW) we do the count of the entire data of all the 5 textbooks as shown in Table 5.4.
- The data analysis for only the meronymic inclusion relation is shown in Table 5.5.
- The count for each dimension – meronymy, class inclusion, spatial inclusion, function, process, and causal is done for all five levels of textbooks for both the LW and RLWs.
- The count for each dimension of linking words is shown in Table 5.6 and the count for each dimension of re-represented linking words is shown in Table 5.7.

We now illustrate the outcome and findings of the data analysis. (Note: The lines (or smooth lines) connecting data points on the graphs shown in Figures 5.12, 5.14, 5.15, 5.17, 5.18, 5.19, 5.20, 5.22 and 5.24 are just a guide to the eye, and do not indicate any functional dependence).

5.4 Outcomes of Analysis

As our focus of the thesis is on the nature and kinds of linking words, we analyzed the data at first for the verbatim usage followed by re-representation. We showcase the patterns that have emerged from the data analysis from the entire data.

5.4.1 Six Kinds of Categories for Relations

The content analysis of the domain has resulted in six categories of semantic relations (Figure 5.2) that are sufficient to represent the entire domain at various levels, from school to college to undergraduate. These are: *meronymic inclusion, class inclusion, spatial inclusion, function, process, and causality.*

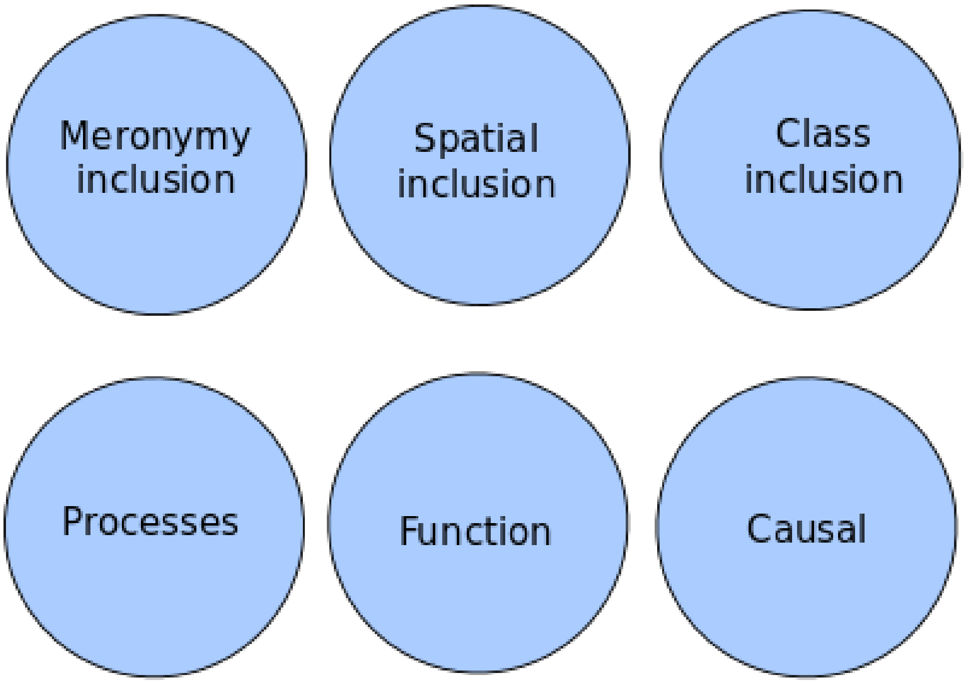


Figure 5.2: Six categories of relations.

5.4.2 Distribution of Relations

The content analysis of the verbatim usage of relations depicts a more significant number of function categories of relations followed by spatial inclusion as shown in Figure 5.3.

During the re-representation of the linking words, the content analysis has resulted in changes in the distribution of the categories. The function category of linking words is reduced as a result of nominalization (see Chapter 3). The process category of linking words along with the causality category of linking words, is increased to a more significant extent. The content analysis of all the levels of text by re-representation has undoubtedly resulted in the process-centric depiction of the text as shown in Figure 5.4. Similarly,

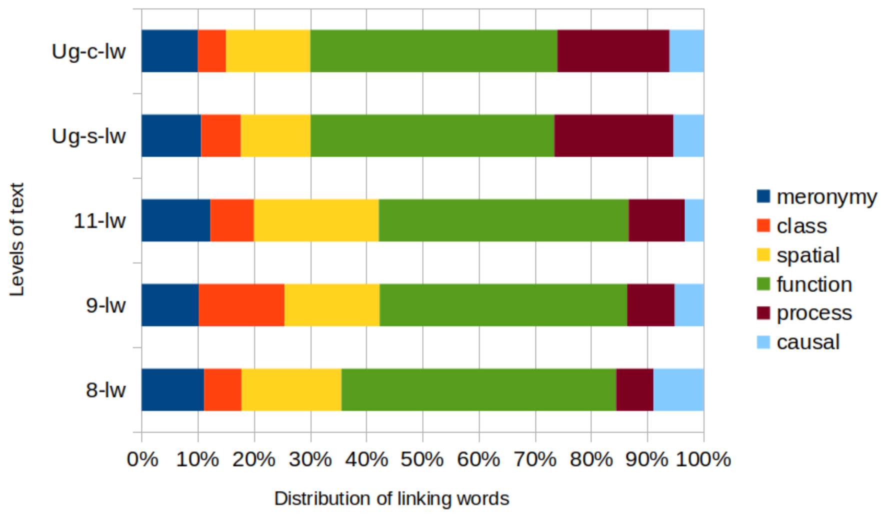


Figure 5.3: Distribution of linking words into six categories.

there has been an increase in the spatial inclusion category as well.

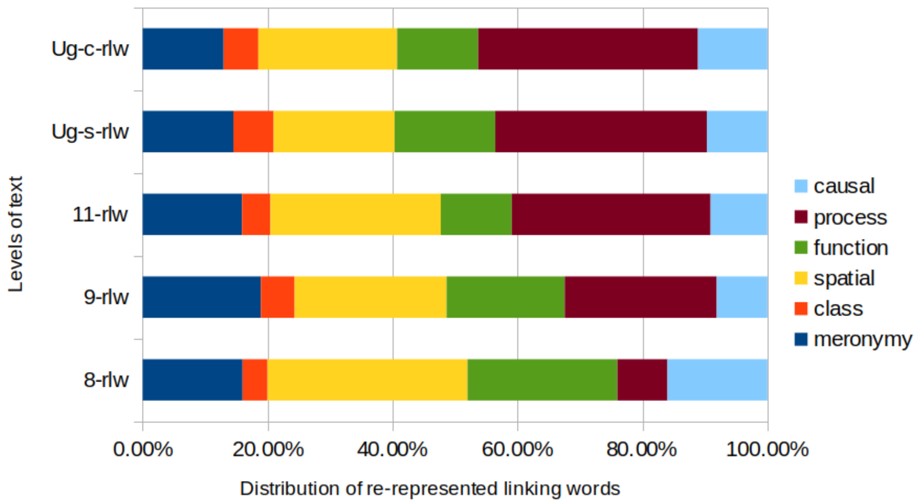


Figure 5.4: Distribution of re-represented linking words into six categories.

5.4.3 Three Kinds of Categories for Attributes

Attribution relation describes a property of an object. The details of making distinctions of relations and attributes are mentioned in Chapter 3. Based on the domain of the content analysis, we categorized the quantitative predicates as attributes, and this resulted in three kinds of categories — *Form, Measurable, Physical Quantity* as shown in Table 5.2. Those propositions that describe the appearance and property, are categorized as Form. The category, measurable, describes any measurable quantity such as size, length, etc. The propositions that describe shapes, color, density, etc., are categorized as Physical Quantity of attributes.

Three categories of attribution (Figure 5.5) have emerged that can represent the entire domain at various levels, from school to college to undergraduate.

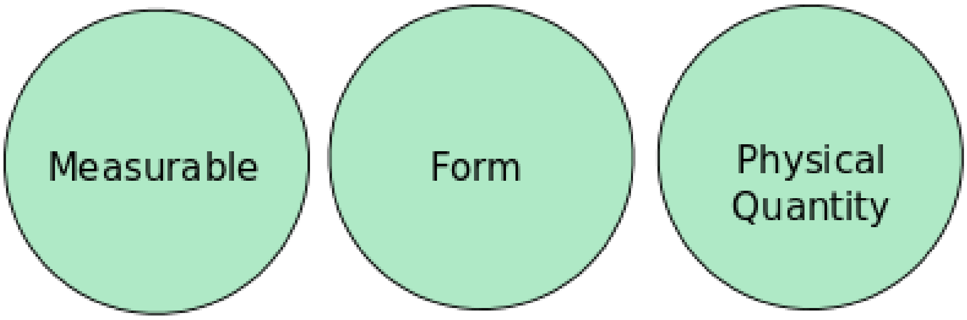


Figure 5.5: Three categories of attributes.

5.4.4 Distribution of Attributes

The content analysis of verbatim usage of attributes results in a larger distribution of the form-category of attributes at the school level, whereas at the college level, it has resulted in an almost similar distribution of Form and Physical Quantity categories of attributes as shown in Figure 5.6.

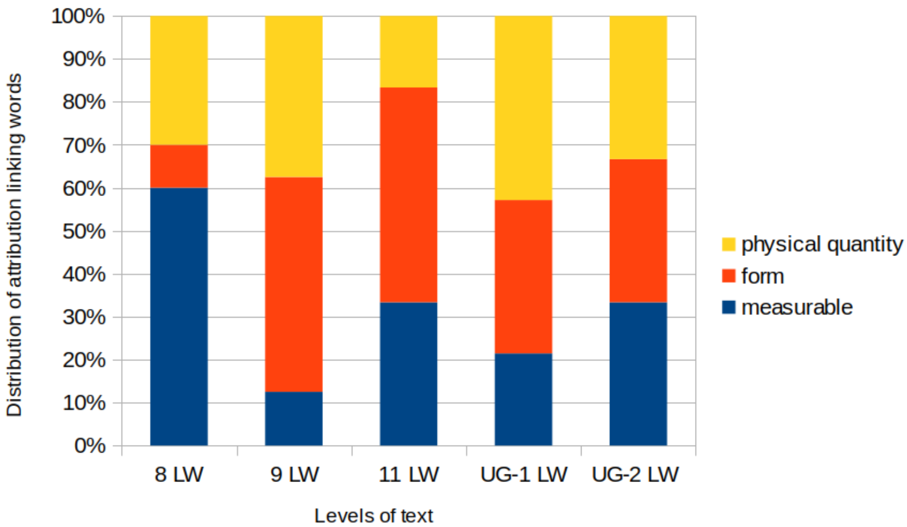


Figure 5.6: Distribution of attributes into three categories.

During the re-representation of the attributes, the content analysis has resulted in changes in the distribution of the categories. In the school-level text, the Form category shows a larger distribution followed by Measurable category, whereas in the college-level text, the content analysis results depict the similar distribution of Form and Measurable categories of attributes as shown in Figure 5.7.

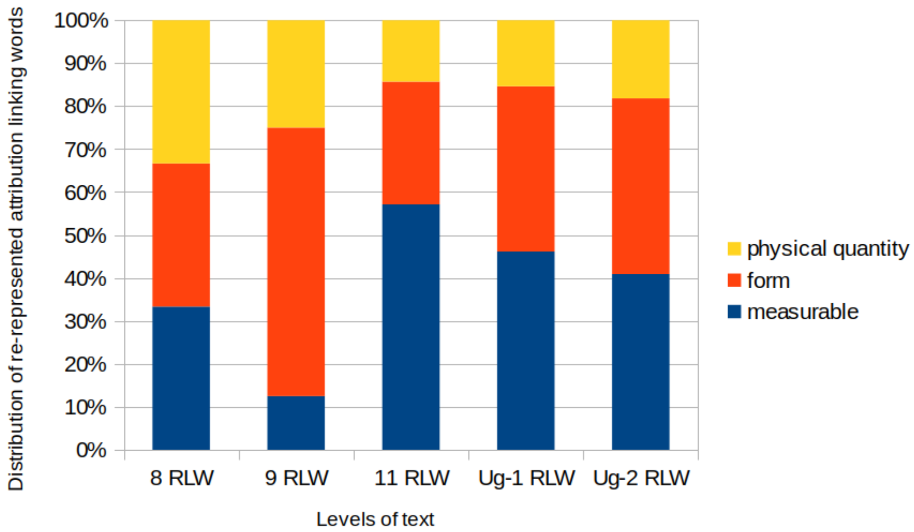


Figure 5.7: Distribution of re-represented attributes into three categories.

5.4.5 Four Ontological Dimensions

Even though the content analysis of all levels of text has resulted in six categories of relations and three categories of attribution, these categories can be grouped into four ontological dimensions. The content analysis emerged into four ontological dimensions of relations that are sufficient to represent the entire cell biology domain. These are depicted in Figure 5.8 and Figure 5.9. The distribution of the verbatim linking words and re-represented linking words into four ontological dimensions of the entire content analysis is shown in Figure 5.10 and Figure 5.11.

5.5 Findings and Patterns

In this exploratory study, the characterization of linking words emerged into specific patterns. We do the characterization on both the paraphrased and re-represented propositions. This analysis helps us to understand whether there is any significant difference between the representations or any change in the nature and number of linking words.

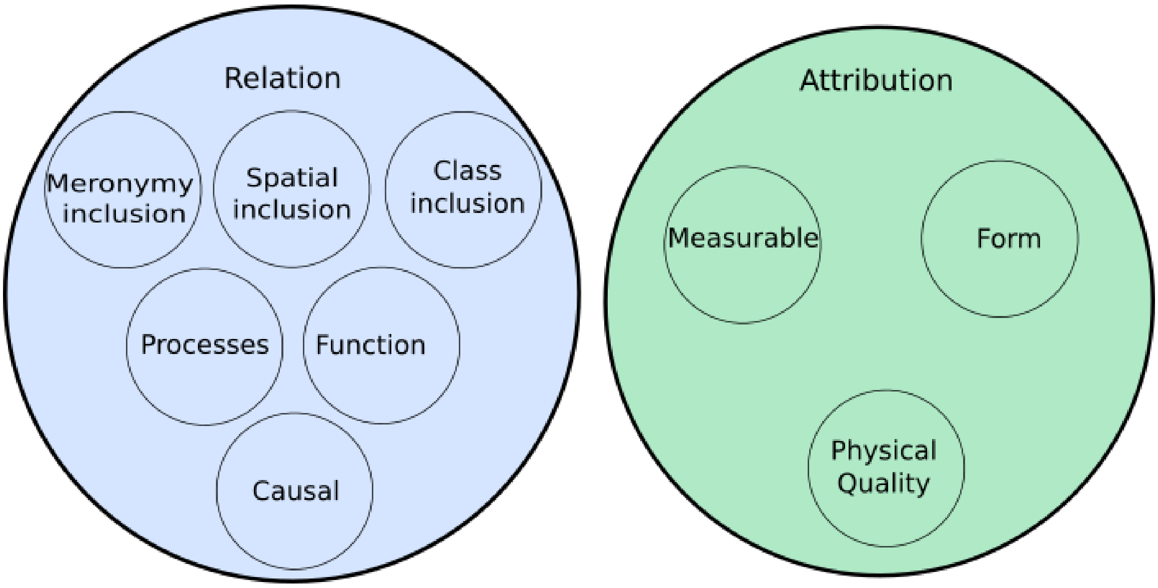


Figure 5.8: All the categories of relations and attributions shown into four ontological dimensions.

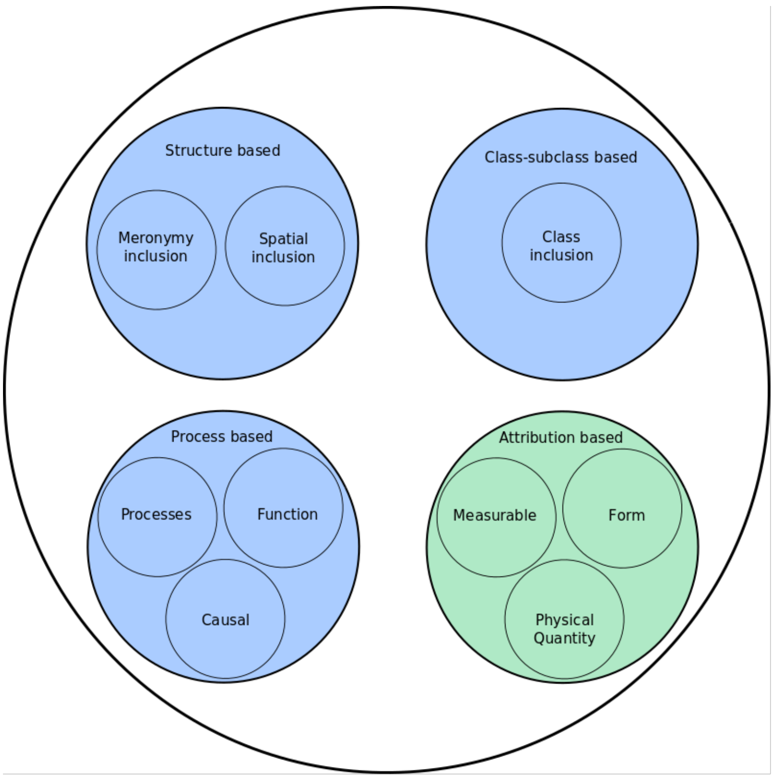


Figure 5.9: All the categories of relations and attributions shown into four ontological dimensions.

5.5.1 Patterns of Verbatim Propositions

The patterns that emerged from the findings of the semantic content analysis of verbatim propositions are:

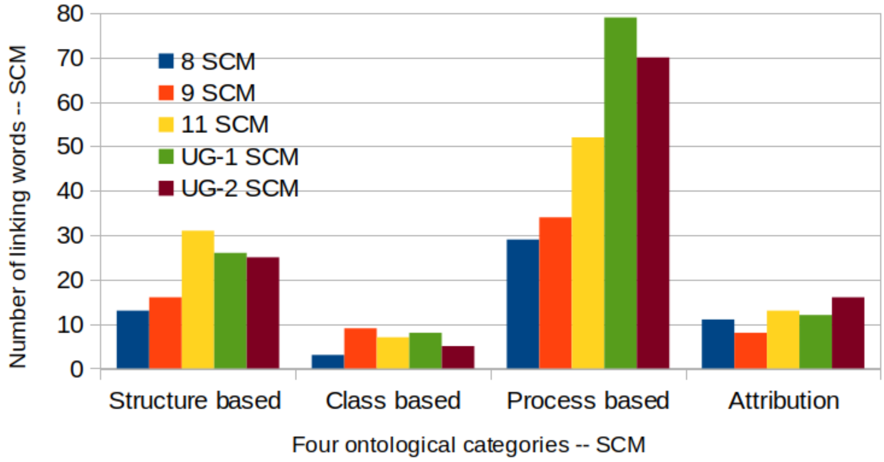


Figure 5.10: Distribution of linking words of the entire content analysis shown into four ontological categories.

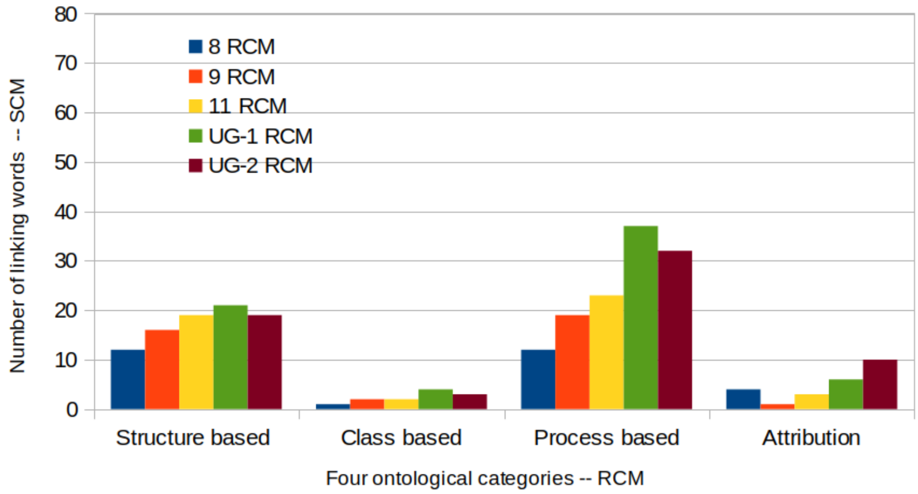


Figure 5.11: Distribution of re-represented liking words of the entire content analysis shown into four ontological categories.

- Increase in concepts, but saturation in linking words
- Prominence of function category of linking words
- Increase in process-based linking words in college-level texts
- Prominence of form category of attribution

5.5.2 Concepts Increase But Linking Words Saturate

One of the highlights from the content analysis has been to indicate that even though the concepts increase at each text level, the linking words attain saturation (Figure 5.12).

It means that, as the level of complexity increases with increasing grades of text, the domain is represented with new concepts. However, even though the concepts increase, the linking words do not grow at the same rate; rather, they attain saturation.

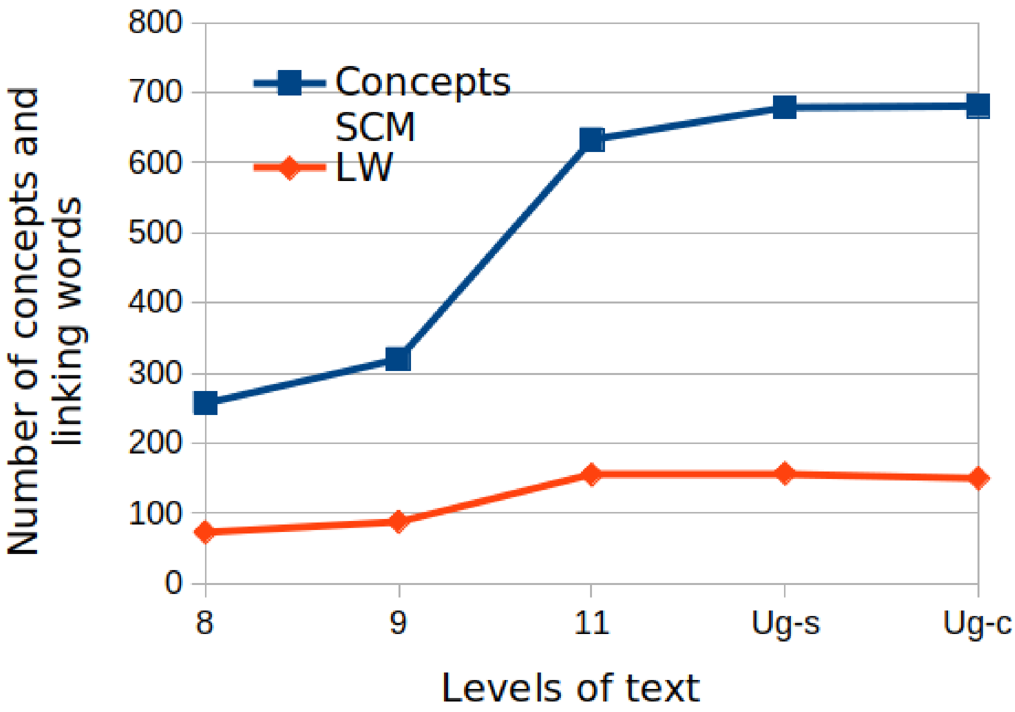


Figure 5.12: Concepts increase with the higher levels of textbooks, but relations attain saturation.

5.5.3 Prominence of Function Category of Linking Words

The result that emerged from the content analysis of the text based on linking words is about the prominence of Function category of linking words as shown in Figure 5.13. This is because most of the text in cell biology has the mention of structure-function relationships. However, it is later evident during the re-representation that this Function category was being used implicitly for every other kind of process as well.

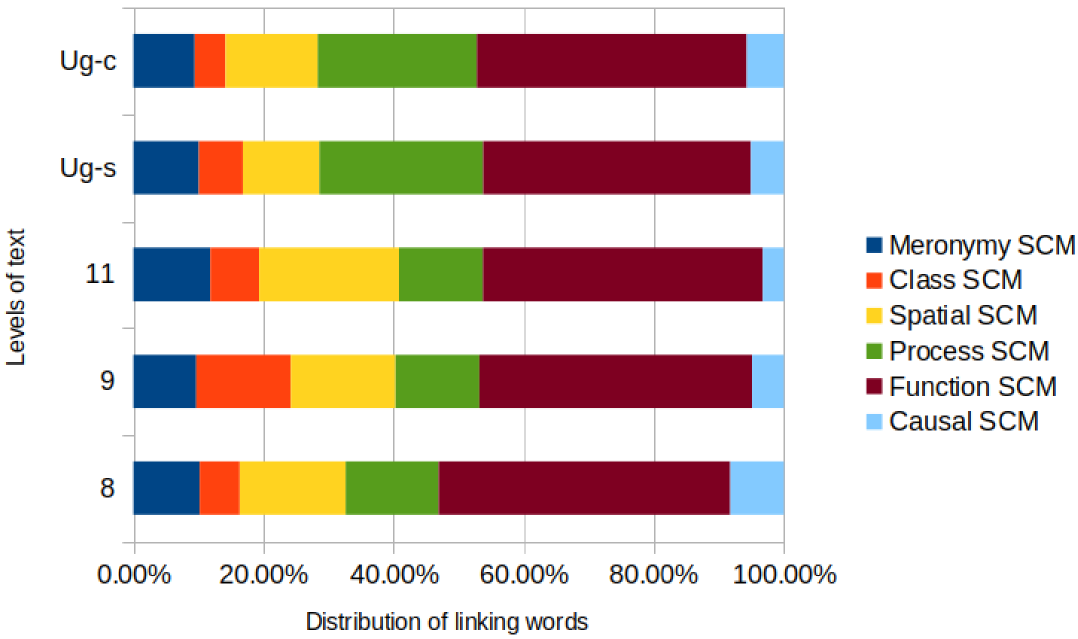


Figure 5.13: Distribution of categories showing prominence of function-based category.

5.5.4 Increase in Process Based Linking Words in College Level Texts

Processes are an essential aspect in the biology domain. The results from content analysis of all the five levels of the text indicates a gradual increase in the process category of linking words as shown in Figure 5.14.

We also conducted the content analysis based on the four ontological dimensions, as mentioned in the previous section 5.4.5. Even then, we get the same results of a gradual increase of process category linking words from school to college level texts as shown in Figure 5.15.

The ‘process’ category is marked by linking words that describe a process, such as moving, passing, becoming, causing, growing, etc. A few examples of linking words from the ‘process’ category are shown in Table 5.8.

During the categorization of linking words across all the levels from school to college to undergraduate, it is observed that the kinds of linking words used to describe processes increase as the topics’ complexity level increases, as shown in Figure 5.14. The ‘process’

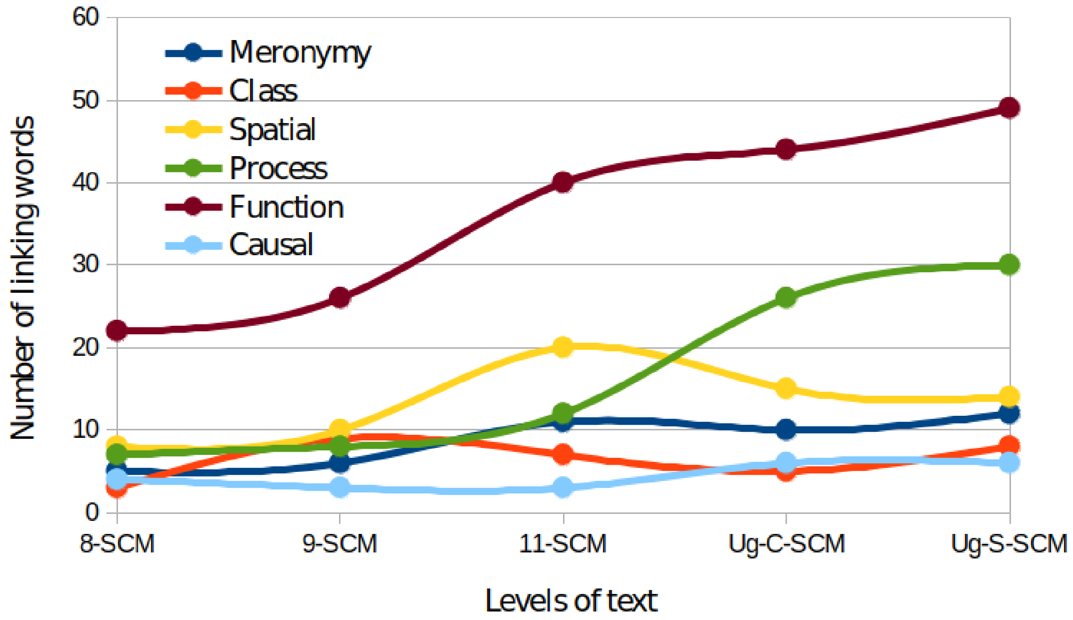


Figure 5.14: Distribution of categories showing higher number of function-based category.

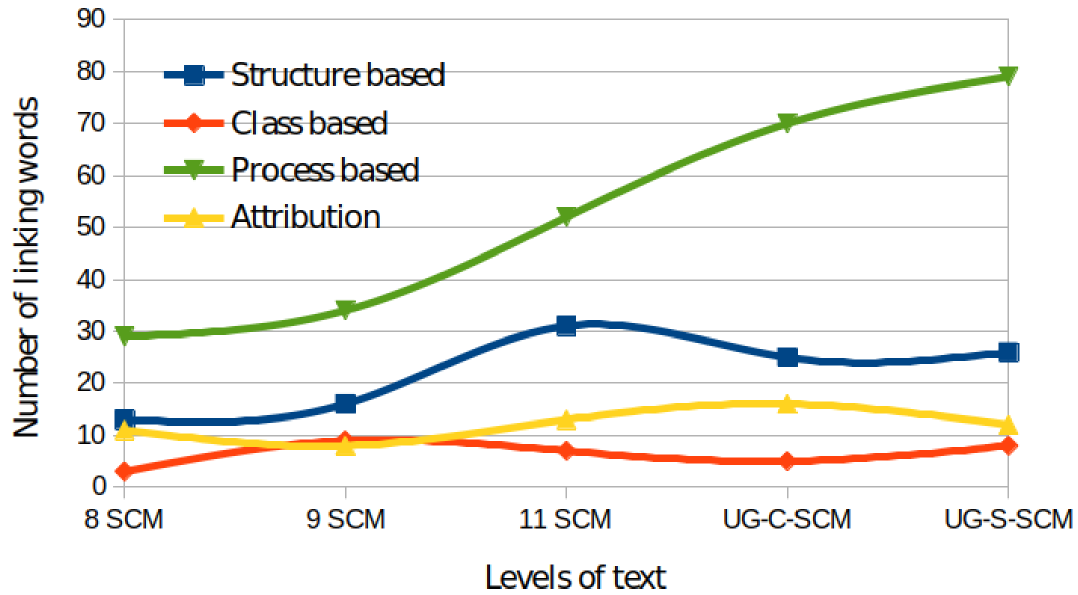


Figure 5.15: Distribution of categories showing higher number of process-based category.

category is increasing with much prominence of two categories – 'developmental' and 'movement' from the high school level onwards. The linking words belonging to the 'developmental' category are 'develops', 'forms', 'bud off into', 'originated', and 'grows'; whereas those belonging to the 'movement' category are 'moves across', 'passes through', 'transports', 'transfers', and 'migrate'. This also indicates that the content becomes more dynamic in description rather than just a static description of the topics. Although the representation domain here is mostly a structure-based, it can also focus on biological processes.

5.5.5 Prominence of Form Category of Attribution

During the content analysis of the attributes, the prominence of the Form category is shown to appear at the school level, with the Measurable category in Class 8. However, there seems to be an almost equal distribution of all three categories of attributes at college level text as shown in Figure 5.16.

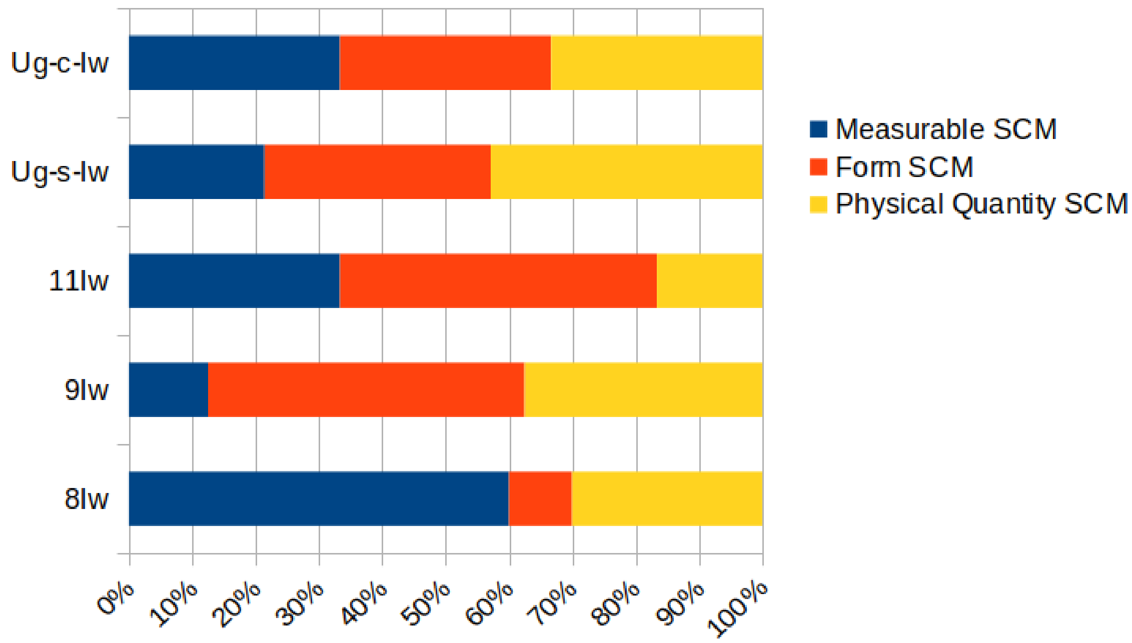


Figure 5.16: Prominence of Form category of Attributes.

5.5.6 Patterns From Re-represented Propositions

We list the following patterns that emerged by re-representation of linking words (RLW):

- Reduced number of re-represented linking words
- Saturation of the ratio between lw and rlw
- Saturation of linking words

- Parsimony in linking words
- Increased number of attribution
- Nominalization of linking words
- Saturation of Function Category of Linking Words
- Process-centric Linking Words

5.5.7 Reduced number of re-represented linking words

Re-representation of the scientific text comprises a significant work in the research. We re-represented all the texts using the re-representation method described in Chapter 3. This resulted in a reduced number of re-represented linking words (RLW) for the same text (Kharatmal and Nagarjuna, 2011), as shown in Figure 5.17.

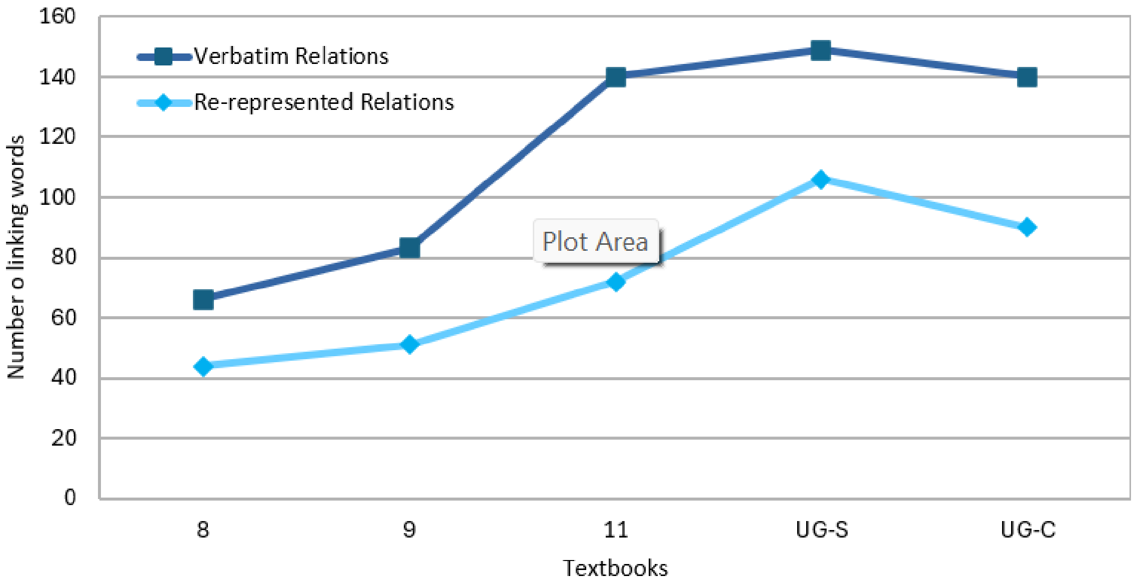


Figure 5.17: Reduced number of re-represented relations.

5.5.8 Saturation of the ratio between LWs and RLWs

An interesting result is that even though the re-represented linking words are reduced at all levels of text, the ratio between the two is almost constant. Figure 5.18 shows the ratio (on the secondary y-axis) attaining saturation or constancy. This indicates that even as the levels of the text increase, the need for new linking words does not increase. The same kind of re-represented linking words are sufficient and can be used even for increasing text complexity. This confirmed and extended our preliminary school data analysis mentioned in Chapter 3.

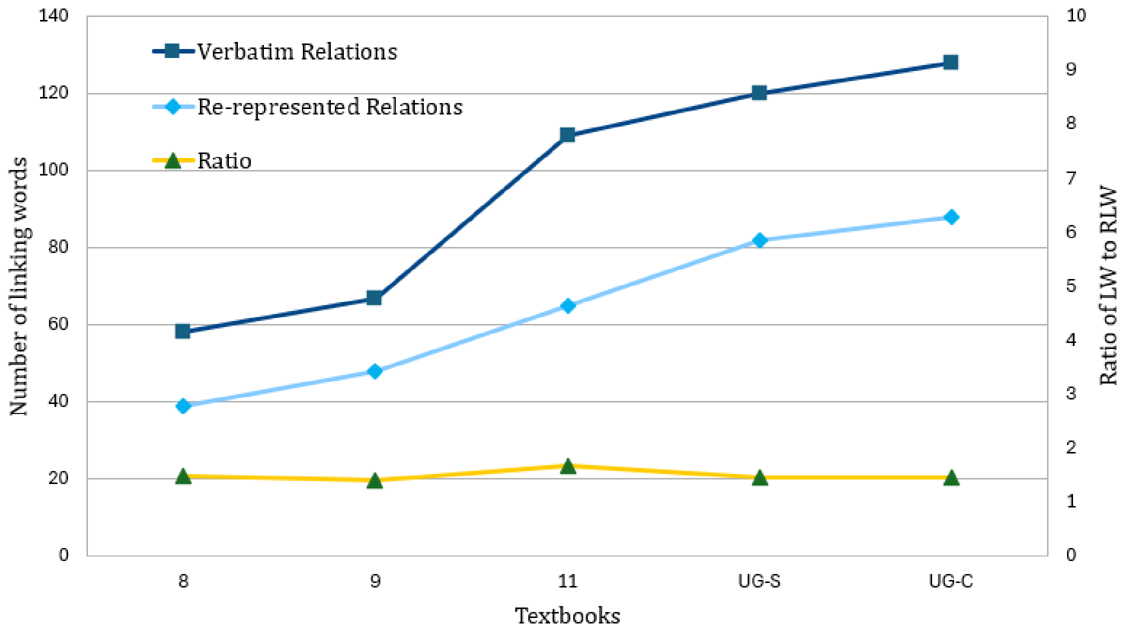


Figure 5.18: Saturation of ratio of verbatim relations and re-represented relations. (Note: Secondary Y-axis depicts the Ratio.)

As the level of complexity increases from school, to college to undergraduate levels, it is expected that as new knowledge is introduced it is by the ways of introducing new terms or concepts. So it is natural that there will be increase in the number of concepts with increasing level of complexity. However, our point is that the requirement of the linking words, that give meaning to this new knowledge or newly introduced concepts, would be the same set, even with increasing level of complexity. At the most the process based relations would show increase from school to college level. To create meaningful propositions, the same set would be used and re-used to maintain consistency. Moreover, the set of linking words is part of everyday language vocabulary and domain-general in contrast to the domain-specific concepts that depend on the topic.

5.5.9 Increase in Number of Attributes

The re-representation of attributes as linking words during the content analysis has resulted in an increased attributes, as shown in Figure 5.19.

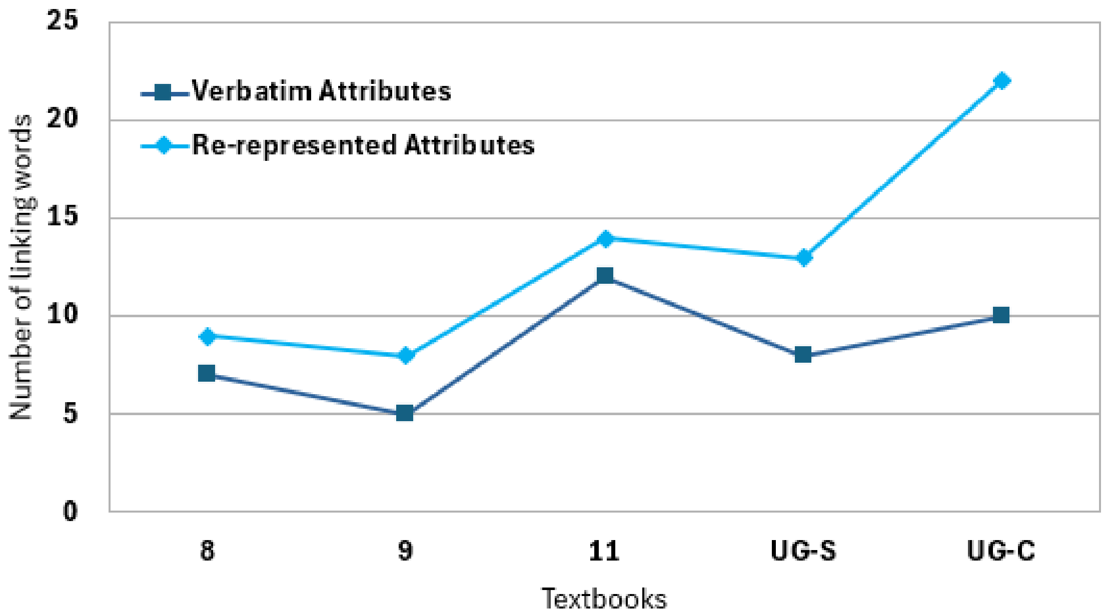


Figure 5.19: Increased number of re-represented attributes as linking words.

An interesting result is that the ratio between the verbatim and the re-represented attributes as linking words is increasing (Kharatmal and Nagarjuna, 2011), as shown in Figure 5.19. It is to be noted that the ratio between relations is achieving constancy (Figure 5.20), but the ratio between attributes is shown to increase (Figure 5.18). This indicates that as the levels of text increase in complexity, so does the need for quantification, and hence the increase in attributes.

5.5.10 Parsimony in linking words

As mentioned above, just as the levels increase, the number of concepts increases progressively. However, the linking words do not grow at the same rate. The linking words attain a saturation which also means not increasing at the same rate as the rate of concepts are growing. This denotes that even when new concepts are introduced, there is no need to coin a new linking word, but the new concepts can be linked with the existing

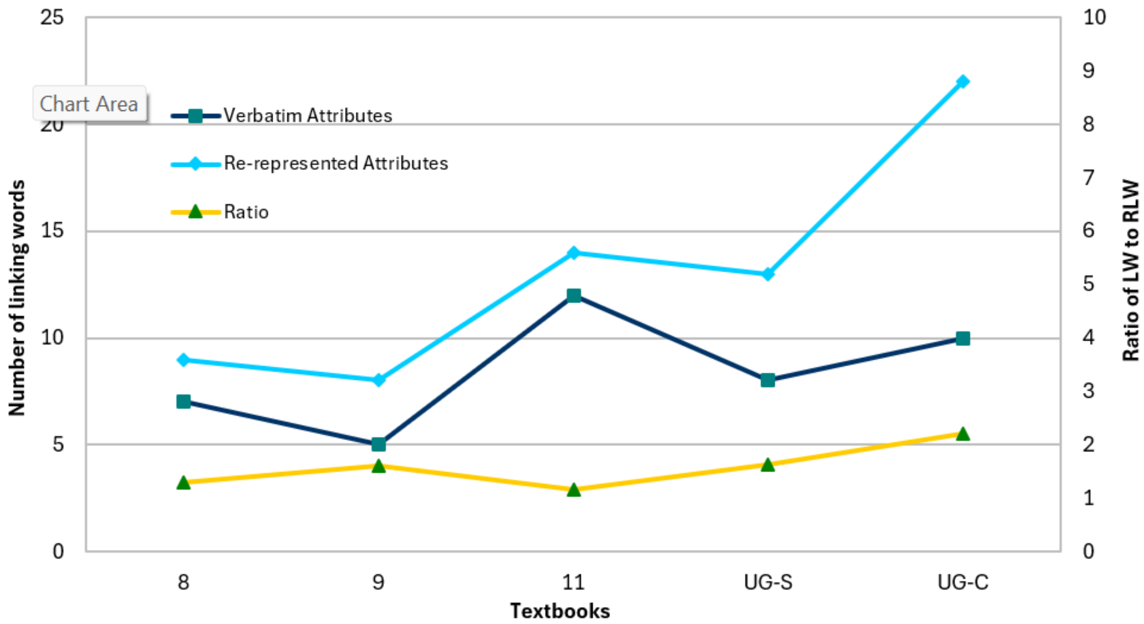


Figure 5.20: Ratio showing an increase in attribution as linking words. (Note: Secondary Y-axis depicts the Ratio.)

vocabulary of linking words.

When we observe the unique set of re-represented linking words, it was found that we can economize on the usage of the re-represented linking words by almost 50% as compared to the verbatim linking words. Thus, re-representation is a method wherein parsimony can be achieved/obtained/maintained by using linking words while creating propositions or concept maps.

5.5.11 Nominalization Leading Towards Process Representation

An outcome of the content analysis is a pattern of nominalization due to the re-representation of text (details of nominalization are discussed in chapter 3). It is known that complexity in scientific language is achieved mainly through terminological density and nominalization of processes and properties. Our work describes and illustrates nominalization in biology text analysis based on Halliday and colleagues' (Halliday and Martin, 1993) framework of Systemic Functional Linguistics (SFL). An advantage of nominalization of a verb or an adjective into a noun is that it takes the form of 'entity', and therefore, it is possible to measure (quantify), generalize, and classify these forms

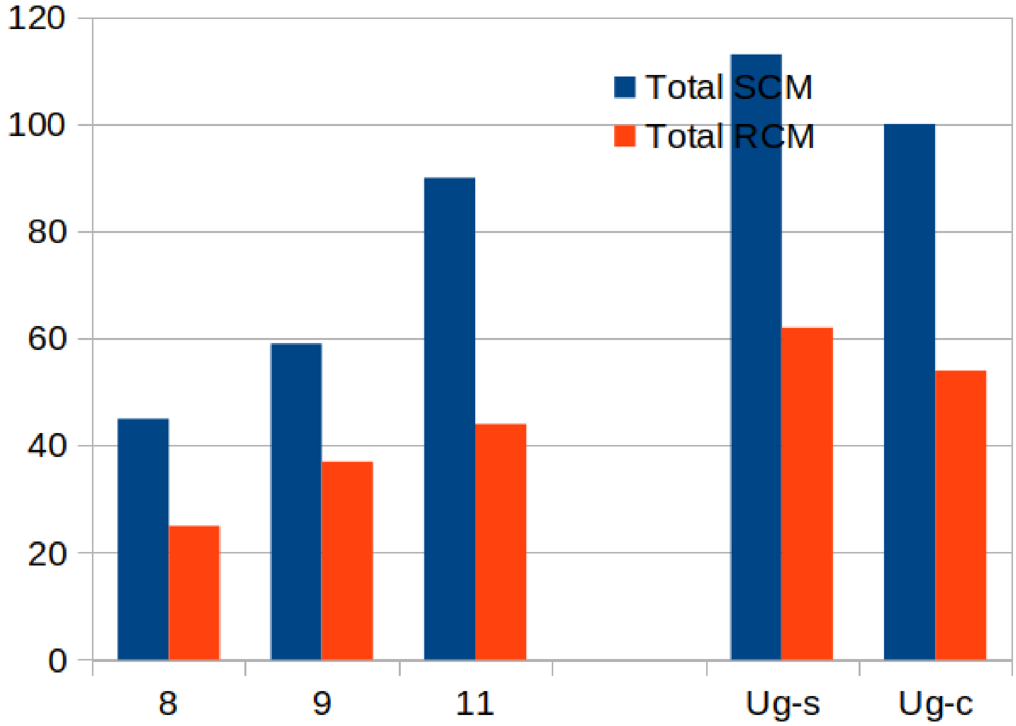


Figure 5.21: Reduction of linking words in re-representation.

(Halliday and Matthiessen, 2004). Halliday explains with the example of “move” taking the form of “motion”, to say, “all motion is relative to some fixed point” leading towards framing laws of motion, discussing perpetual motion. This occurs by the transformation of “motion” from being just a thing into a phenomenon. Further, Halliday says that by “just calling ‘move’ motion, we have not changed anything in the real world; but we have changed the nature of our experience of the world” (Halliday and Matthiessen, 2004, p. 16).

5.5.12 Saturation of Function Category of Linking Words

We illustrate a collated verbatim list of 57 linking words used in the school texts of classes 8, 9, 11, categorized as Function. As a result of re-representation, these are reduced to just 10 kinds of linking words used for the same Function category as shown in Table 5.11.

Not only does the nominalization highlight the process-oriented nature, but it also economizes the use of linking words. We can see that while re-representing action verbs, we use only very few kinds of linking words ("has function", "has role", etc.) as shown in Table 5.11. This results in a reduced number of Function categories of linking words, as shown in Figure 5.22. This method of nominalization and the data analysis was approved by the SFL community (Kharatmal, 2024).

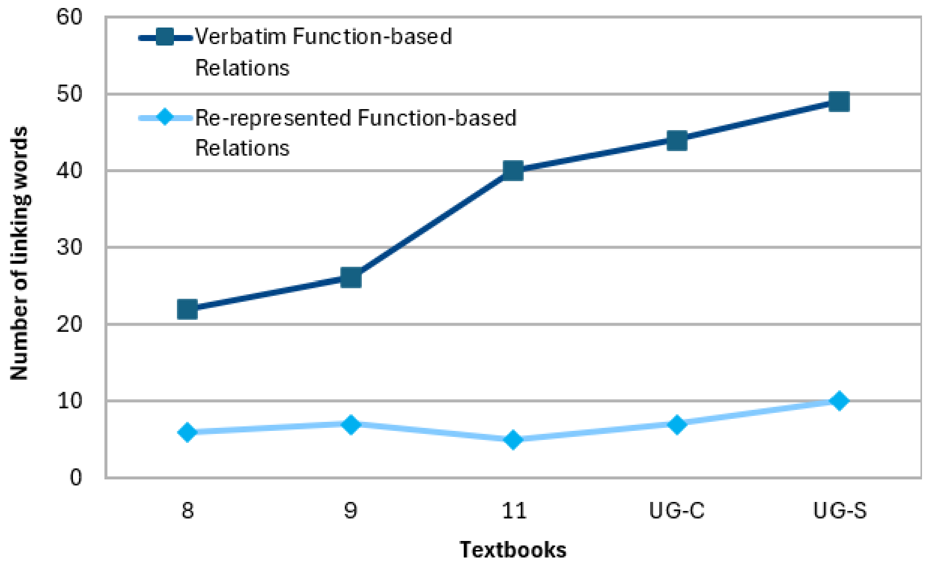


Figure 5.22: Saturation of function-based relations.

During the nominalization of the Function category, the linking words are assigned as process nouns. This re-representation helps in rendering the biology text towards a more process-centric approach than a structure-centric.

In the verbatim text, there were process-related words that required nominalization, but while doing re-representation, these process-based linking words are nominalized. The content analysis of the text shows an increase in the non-nominalized linking words as the content levels increase; therefore, the need for nominalization increases. Once these linking words are nominalized, it attains a saturation of not needing more than four kinds of re-represented linking words (Table 5.12). Figure 5.23 indicates a reduced number of function-linking-words and an increased number of process-linking-words due to differentiation, from the content analysis of the school-level texts.

This results in understanding the role/function of any structural concept in a process. The re-representation resulting in the nominalization of the process has a significant implication. The text, otherwise considered to be a description of the structural aspect of cell biology, brings the importance of physiological and molecular processes into the foreground. The re-representation method has been worked out to model the processes as well.

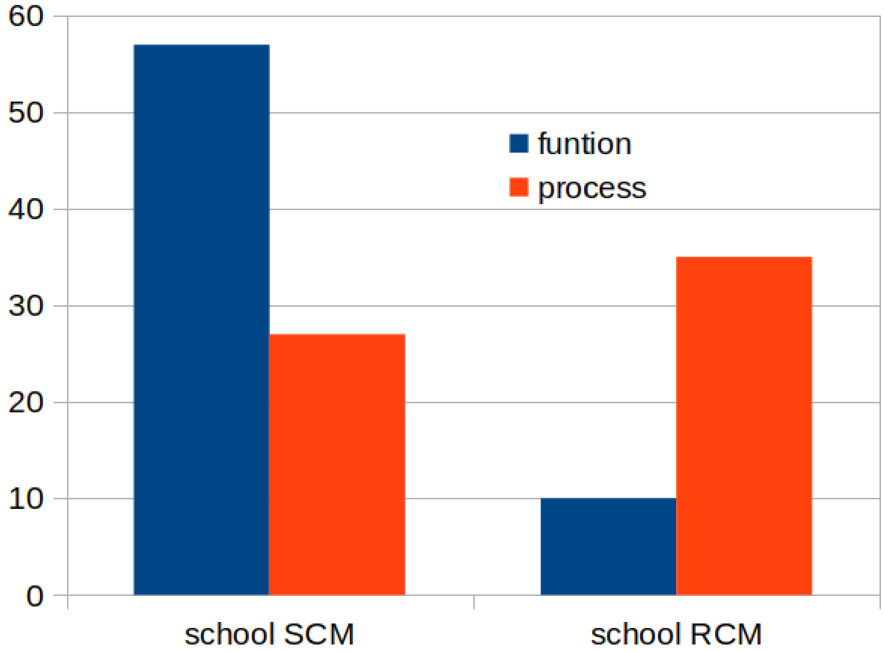


Figure 5.23: Data showing the increase in process-based linking words due to nominalization of Function category.

5.5.13 Disambiguation of Process Based Linking Words

During the content analysis, the verbatim proposition is re-represented only when the linking words are ambiguous. However, we observed that the process-based linking words are almost similar in both i.e., there was no need for re-representation because the verbatim linking words have semantically accurate usage of linking words as shown in Figure 5.24.

This highlights that ambiguity is not a concern/factor when describing processes. This is in contrast with observations of ambiguity appearing in describing structure (meronymy inclusion, spatial inclusion), describing classification (class inclusion), and describing property (attribution). Further investigation may be required to find out if textbooks use identical words for processes across different levels.

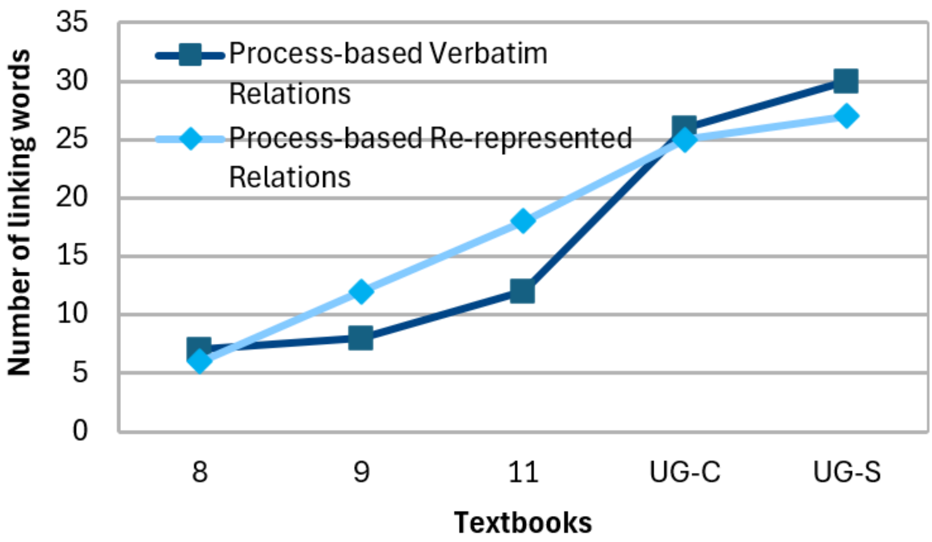


Figure 5.24: Comparison of process-based relations in verbatim and re-representation methods.

This feature also helps to realize that even though the process-based linking words belong to the natural language vocabulary, they still provide accurate sense while describing the processes in a proposition. This may be because these linking words are connected to either structure and process or process and process. The subject and object places are certain/fixed, and therefore, there is no scope for ambiguity.

5.5.14 Process-centric Re-represented Linking Words

An interesting observation based on the re-representation of text is that the Function category that appears in the verbatim linking words is reduced using the re-representation method. In addition, most of the 'function' category linking words are re-represented by converting these into process concepts. Thus, a static entity becomes a process-oriented entity depicting biological processes, an essential aspect in understanding biology.

We put forward that it is only after the processes are in the subject place and become focused, that we can model these processes in terms of changes in state, movement, development, growth, transformation, etc. The nominalization process not only brings the process into focus but also results in a specific and explicit nature of process-oriented linking words. An essential outcome as a result of highlighting the processes enabled to specific sub-categories – *process, change and development, movement, temporal, causal,* *and function.*

We mention that the nature of process-based biology texts has a finite set of categories. A list of each of these kinds of linking words is – 'Processes' as shown in Table 5.13; 'Change and development' as shown in Table 5.14; 'Movement' as shown in Table 5.15; 'Temporal' as shown in Table 5.16 and 'Causal' as shown in Table 5.17. The Function category is already listed in Table 5.11.

All school biology texts in the context can be categorized into these five sub-categories. We show a comparison of the distribution of process and function category of linking words in a verbatim and re-represented method in Figure 5.25 for school text and in Figure 5.26 for college text. In both the figures depicting the results of re-presentation, there is an increase in the function category of linking words. This has led to distribution and differentiation of process-based categories of linking words, giving a greater scope for a process-centric approach.

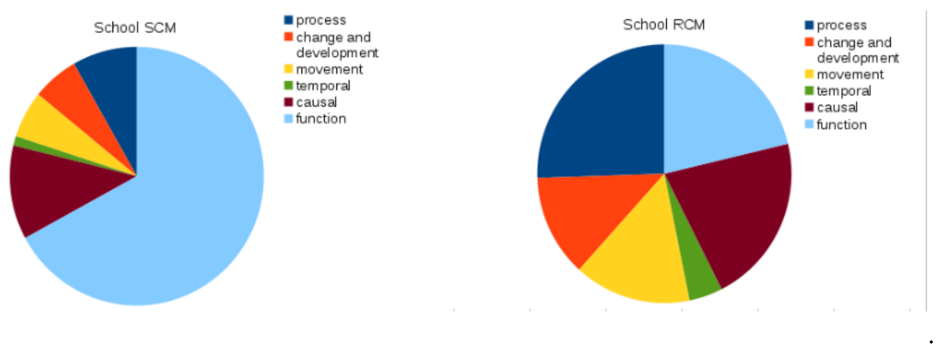


Figure 5.25: A comparison of the distribution of process and function category of linking words in verbatim and re-representation of school text.

5.6 Characterization for Knowledge Restructuring

We wanted to explore what kinds of restructuring can occur due to re-representation. Restructuring or conceptual changes in the context of the history of science are categorized as – *addition, deletion, category shift, coalescence, differentiation* (Thagard, 1992). While the conceptual change research focuses on the reorganization of concepts, we observed

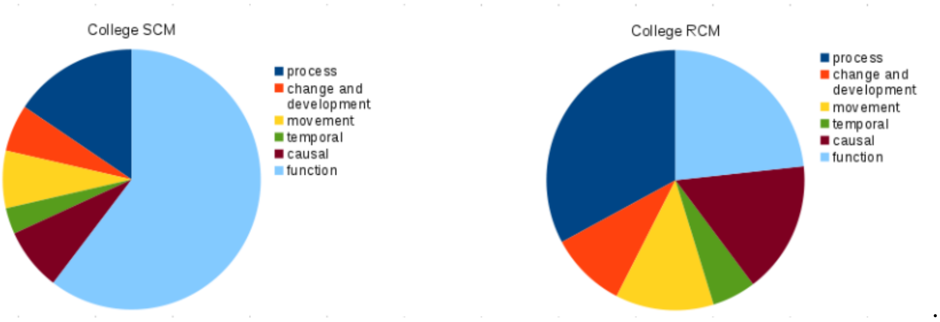


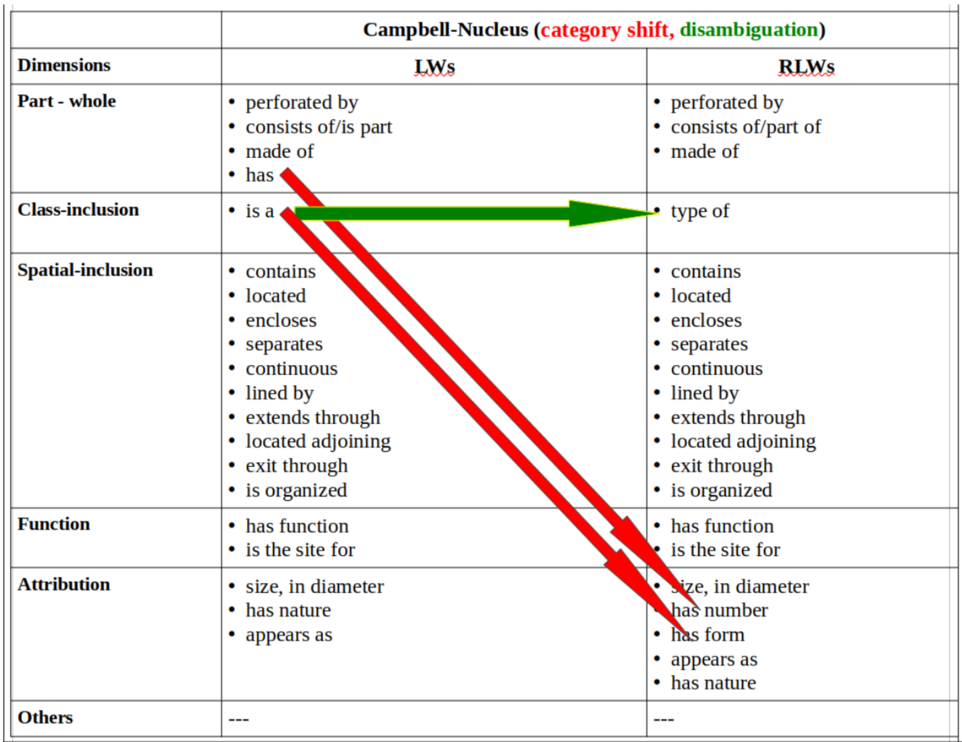
Figure 5.26: A comparison of the distribution of process and function category of linking words in verbatim and re-representation of college text.

these kinds of restructuring during the re-representation of linking words. We elaborate with illustrations from the content analysis.

5.6.1 Category Shift

Category shift (Vosniadou, 2008), or reclassification or branch jumping (Thagard, 1992) can be depicted by ontological miscategorizations of concepts as a shift across lateral categories (Chi and Roscoe, 2002). Figure 5.27 shows category shift and disambiguation in the RLWs.

In C-N, the LW 'is-a', which denotes a class-inclusion dimension that has been used to connote the attribution dimension. Not only is category shift involved, but also disambiguation occurs in the RLW. In the propositions, 'nuclear membrane is a double membrane'; 'each membrane is a lipid bilayer', the first one would become more explicit and clear when its RLW is changed to "type of", while the latter one since it describes a form i.e., bilayer, it would be more appropriate to change it to RLW "has form". Due to category shift, 'is-a' has been changed to "has form". and on the other hand, due to disambiguation, "is-a" has been re-represented as "type of". In the part-whole dimension, the LW "has" is changed to the attribution dimension as "has a number". The proposition 'human cell has 46 chromosomes'; 'human sex cells have 23 chromosomes' display the part-whole dimension due to the usage of "has", and "have". In order for an explicit re-representation, the more appropriate RLW would be the usage of "has number" and therefore belongs to the attribution dimension. This restructuring from part-whole to



	Campbell-Nucleus (category shift, disambiguation)	
Dimensions	LWs	RLWs
Part - whole	• perforated by • consists of/is part • made of • has	• perforated by • consists of/part of • made of
Class-inclusion	• is a	• type of
Spatial-inclusion	• contains • located • encloses • separates • continuous • lined by • extends through • located adjoining • exit through • is organized	• contains • located • encloses • separates • continuous • lined by • extends through • located adjoining • exit through • is organized
Function	• has function • is the site for	• has function • is the site for
Attribution	• size, in diameter • has nature • appears as	• size, in diameter • has number • has form • appears as • has nature
Others	---	---

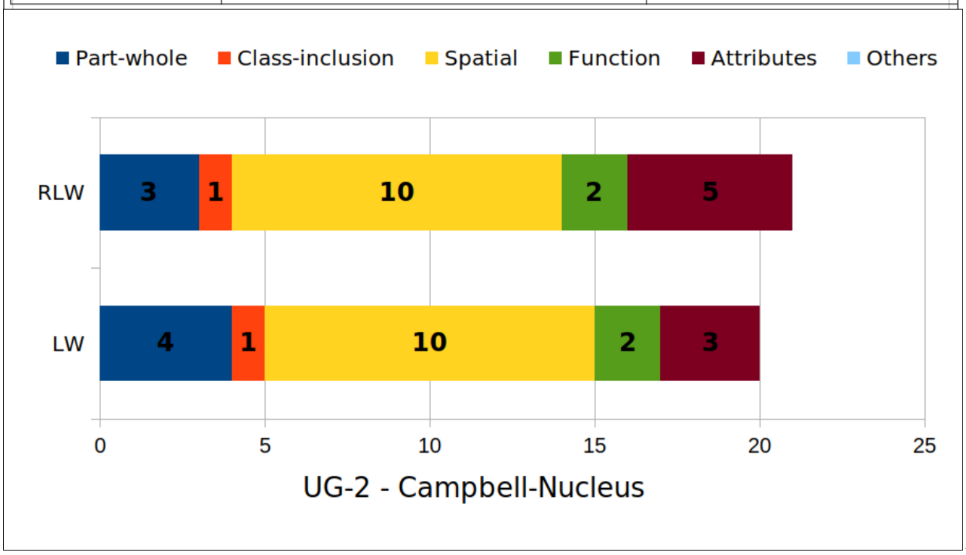


Figure 5.27: The analysis from Campbell-Nucleus (C-N) depicting category shift, disambiguation in the RLWs.

attribution is a category shift.

5.6.2 Coalescence

In the DeRobertis-Mitochondria topic (D-M), although there is a decrease in the overall number of RLWs, it can be seen that within the RLWs there is a decrease in the class-inclusion dimension. At the same time, there is an increase in the attribution dimension. In the propositions, 'mitochondrial crests are continuous, are fluid, is gel-like', the LWs "are", "is", "like", usually which belong to the class-inclusion relations, actually describe the properties, and so these LWs are re-represented as "has property". This change from one dimension to the other i.e., from class inclusion to attribution, results in a category shift. Thus, an attribution relation is added in RLW. Within the spatial-inclusion dimension, the LWs "filled with" and "contains" are coalesced into "contains" in the RLW. This helps in maintaining disambiguity and consistency in the RLWs.

The LWs in the spatial-inclusion dimension, "contains", and "with" are coalesced into "contains" in the RLW, resulting in the decrease of the relations. The propositions 'inner membrane contains matrix'; 'matrix with ribosomes' can be represented with the linking word "contains" alone. This helps to achieve parsimony in the use of relations.

The LWs in the spatial-inclusion dimension, "contains", and "with" are coalesced into "contains" in the RLW, decreasing relations. The propositions 'inner membrane contains matrix'; 'matrix with ribosomes' can be represented with the linking word "contains" alone. This helps to achieve parsimony in the use of relations (Figure 5.28).

5.6.3 Differentiation

In the topic of Soper-Nucleus (S-N), "are" is used for class-inclusion and attribution in the LWs, thereby causing ambiguity in the propositions "two nuclei are micronucleus, meganucleus"; "nuclei are 10 mm in diameter". The same linking word is disambiguated and re-represented as "kind of" in the former proposition and "has size" in the latter proposition. This helps to provide semantic accuracy. It can also be seen that even within the attribution dimension, "is" and "are" are used, which is made more explicit when they get differentiated into "has property" and "has diameter". This is shown in Figure 5.29.

5.7 Summary

In this chapter, we addressed research question 2. We presented the findings in terms of outcomes and patterns as a result of using the re-representation method for the semantic

	DeRobertis-Mitochondria (coalescence, category shift)	
Dimensions	**LWs**	**RLWs**
Part - whole	• consists of	• consists of
Class-inclusion	• divides into • are, is • like	• divides into
Spatial-inclusion	• surrounds • projects into • contained between • filled with • contains	• surrounds • projects into • located between • contains
Function	---	---
Attribution	• has nature • has size	• has nature • has shape • has size
Others	• called	• called

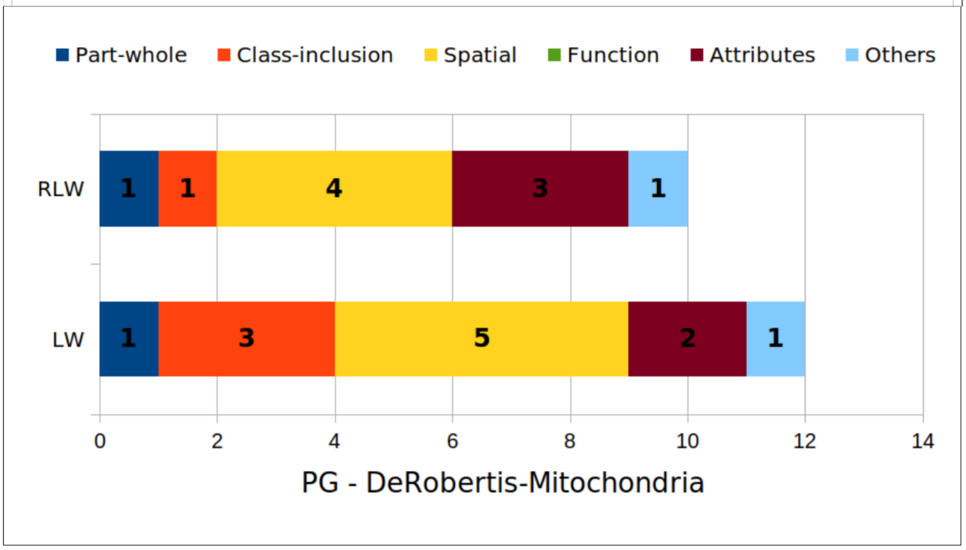


Figure 5.28: DeRobertis-Mitochondria (D-M) analysis depicting category shift and coalescence in the RLWs.

content analysis of the text.

Some of the significant findings from the study resulted in emerging of six kinds of categories for relations and three kinds of categories for attributes. This culminated into four ontological dimensions – structure-based, class-subclass-based, process-based and attributes – that are sufficient for representing cell biology.

An important finding of the study has resulted into saturation of linking words in the growth of knowledge. On the other hand, there is increase in attributes depicting increasing quantification at higher levels of biology textbooks, and increase in process-based linking words depicting dynamic propositions. The nominalization has resulted into

	Soper-Nucleus (disambiguous, differentiation)	
Dimensions	**LWs**	**RLWs**
Part - whole	• exists • composed of • may be present	• exists • composed of • may be present
Class-inclusion	• such as • is, are	• for example • type of, kind of
Spatial-inclusion	• found • in the form of • surrounded by • is around • covered with • bound to • wound around • packed in • occurs near • are inside	• contains • in the form of • surrounded by • is around • covered with • bound to • wound around • packed in • occurs near • located inside
Function	• has function • formed by	• has function • formed by
Attribution	• are, is • are largest • are 10mm • seen as • appears as	• has property • has size, large • has length, 1m • has diameter, 10mm • has form, dark patch • has appearance, rounded
Others	• is continuous with • called • means • is	• is continuous with • called • means • is

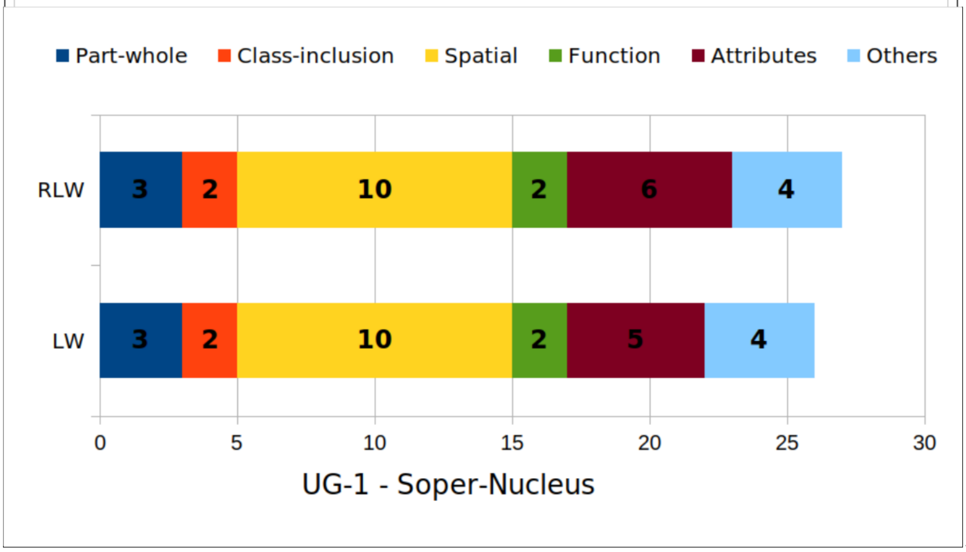


Figure 5.29: The analysis from Soper-Nucleus (S-N)) depicting differentiation in the RLWs.

saturation of Function category and differentiation of processes. The re-representation of linking words can also be considered for knowledge restructuring.

We try to demonstrate the potential of the developed method for content analysis i.e. specifically focusing on semantics. We agree that any analysis could emerge into certain outcomes and patterns. This would also depend on what kind of questions we are asking for analysis. Here we would like to clarify that, we wanted to bring the focus on and make its usage explicit. Through this we have been able to highlight the process-centric nature of biology which was implicit in the text. This could emerge when we disambiguate the function based linking words and nominalize the verbs for processes. We also wanted to highlight the difference in the nature of knowledge before and after re-representation, that lead to reduced number of linking words, giving scope for making these distinct from attributes. Through this we have tried to unravel the nature of biology knowledge.

However, we also provide some thoughts on in what context would the method provide negative results. We think, perhaps for the representations of physics, mathematics, chemical equations, graphs, scientific nomenclature (IUPAC), periodic table, scientific laws, axioms, rules, classification system. Because these apply the principles of rigor, disambiguation for expression in language.

8 lw	9 lw	11 lw	UG-1 lw	UG-2 lw
absent in/in/present in/ scattered in	absent in/lack/present in/ present only in/ present as part of/in	absent in/are present in/has the presence of/lack	are /are in/is a	absent in/lack
common in	are full of/filled with/embedded in/	are/are in/in	associated with/joined together to form/remains together/accumulates in/breaks up into/cement together/deposited in	add/add to its/built into/constructed from
has/has no	consists of/constitute	consists of/has/has central part/have/have their own	consists of/constitutes	consists of/part of/portions of/of
together constitute/are components of	does not have/has/have	distributed among/scattered in/stored in/lie /bear/show/filled with	fills with /stored as	embedded in/fuses with/deposited /perforated by/anchored by
	gets organized into/groups together into/is associated with/not with	found in/not found in	group of/of the //form	encompasses/organizes
		interrupted by/fuse with/associated with/together form	have/have same	has
		is/is of/is mainly of/is with/with	is continuous with/has partial/extending through/is perforated by	involved in/in /work together with
		organized as/organized into/form/formed from	is inside/is in/is packed in/inside	organizes
		possess/bear	possess	
			present in/shows presence of/lack/exists in	

Table 5.3: Representational data of Meronymy inclusion dimension of verbatim linking words (LW).

8 rlw	9 rlw	11 rlw	UG-1 rlw	UG-2 rlw
absent in/present in	associated with/ form into/ organized into	associated with/fuse with/interconnected with/c̃oordinated with	consists of/constitutes	assembled with/in/combines with/constructed from
comprises of	consists of/not consists of/part of/constitute	consists of/p̃art of/c̃onsists of/part of/constitute	forms/deposited in/fills with /impregnated with/stored as	consists of/does not consist of/ part of/not part of
consists of/part of/together constitute	embedded by	embedded in	is continuous with/is perforated by/are extending through/	embedded in/present in
	continuous with	organized into/form	is packed in/joined together to form/associated with/accumulates in	organized into
	present in/lacks	present in	possess	
			present in/ show presence of/inside/lack/involved in	

Table 5.4: Representational data of Meronymy inclusion dimension of re-represented linking words (RLW).

Levels	Total LW in Meronymy	Total RLW in Meronymy
8	5	4
9	6	7
11	11	7
UG -1	12	9
UG -2	10	8

Table 5.5: Total count for different kinds of meronymy relations. The Total LW and RLW for each level is tabulated as shown in Table 5.3 and 5.4, respectively.

Dimensions of linking words (LWs)	8	9	11	UG-1	UG-2
Meronymic inclusion LW	5	6	11	12	10
Class inclusion LW	3	9	7	8	5
Spatial inclusion LW	8	10	20	14	15
Function LW	22	26	40	49	44
Process LW	7	8	12	30	26
Causal LW	4	3	3	6	6
Total	49	62	93	119	106

Table 5.6: Dimension-wise total count of kinds of linking words of all the dimensions at all the five levels of the textbooks of entire data.

Dimensions of re-represented linking words (RLWs)	8	9	11	UG-1	UG-2
Meronymic inclusion RLW	4	7	7	9	8
Class inclusion RLW	1	2	2	4	3
Spatial inclusion RLW	8	9	12	12	12
Function RLW	6	7	5	10	7
Process RLW	6	12	18	27	25
Causal RLW	4	3	4	6	6
Total	29	40	48	68	60

Table 5.7: Dimension-wise total count of kinds of re-represented linking words of all the dimensions at all the five levels of the textbooks of entire data.

Process Categories	Concepts	Process linking words	Concepts
Process	Cell	undergoes process	cell division
Movement	Substances	move across	cell membrane
Process	Cleaning of cell	occurs by digesting of foreign material	
Movement	Digestive enzymes of pancreas	passed to	duodenum
Development	embryo	develops into	chick
Causal	Propelling microfilaments	causes	contraction of muscle cells

Table 5.8: Process-based linking words.

Levels	Concepts	LW	RLW
8	257	45	25
9	320	59	37
11	633	90	44
UG-1	679	113	62
UG-2	681	100	54

Table 5.9: Number of concepts and LW and RLW of all the levels.

Levels	Total LW	Total RLW	%Reduction in RLW	% Extra in LW
8	45	25	55.56	44.44
9	59	37	62.71	37.29
11	90	44	48.89	51.11
UG-1	113	62	54.87	45.13
UG-2	100	54	54.00	46.00

Table 5.10: Reduction of LWs in re-representation at all the levels.

List of linking words used to represent function category	List of re-represented linking words to represent function category
• act as • allow/permit/prevent • are • can/may • change • capable of/enables • carry / carrying outdated • controls • create • determines • digest • divides • do not play role • emerge from • essential for/required for • exposed to • facilitates • formed by/ forms • function as / has function • gives/gives rise to/provides • helps in • holds/glues • impart • important for • inform of • increase • interacts with • involved in • keep • make • modified in • multiplied • need protection against • perform • prevents • protects / protected from • receives • released from / released by • required for • responsible for • revealed • sends the • separates • site for / site of / site of formation of • stores • support • synthesizes • to pump • transfers	• act as • exposed to • formed / forms / required for forming • has function • has role • has tendency to • helps in • site for • synthesized for • use

Class	Number of linking words before nominalization (LW)	Number of linking words before nominalization (RLW)
8	15	2
9	22	4
11	35	3
UG-1	44	4
UG-2	42	4

Table 5.12: Data showing non-nominalized and nominalized linking words.

List of linking words used to represent process category	List of re-represented linking words used to represent process category
• passed, cannot pass • by, enter • looses • through • show • takes place against • condenses into • exits • extracted by • forms • obtained from • occurs • produce • reassemble into • seem to • stains • together form • undergo • works in • divides • eat by • grow from • process of • involves • propel through • recycle half	• passed • enters • exerted against • involved in process • occurs in • undergoes process • passed to/cannot pass through • changes to • released from • shows • breakdowns • condenses into • entered into • extracted by • formed by • join/attaches/ • stopped by • obtained in • occurs in • stains • added to/extend into • has process • interacts with • stops • released from • translates

Table 5.13: List of process-based linking words.

List of linking words used to represent development category	List of re-represented linking words used to development category
• modifies in • begins life • come from • ends up in • develops • thickened to become • becomes • bud off from • develops • grows • pinch into	• modifies • develops • becomes • developed from • formed • thickened to • develops • grows out

Table 5.14: List of development-based linking words.

List of linking words used to represent movement category	List of re-represented linking words used to movement category
• traversed by • delivered to • is passaged through • move / move towards • transported across / within • enters / enters through / by • move from /to • passed into • runs through • travels to • migrate • movement around • transferred	• movement across • through • traversed by • delivered to • is passaged through • move/ move towards • transported • exports through • exits • moves along • runs through • reaches

Table 5.15: List of movement-based linking words.

List of linking words used to represent temporal category	List of re-represented linking words used to temporal category
• when • followed by • laid down during • were once • during	• before • during • followed by • precedes • were once

Table 5.16: List of temporal-based linking words.

List of linking words used to represent causal category	List of re-represented linking words used to causal category
• because • does not affect • results into • led to • causes • therefore • can stretch • depends on	• because • does not affect • due to • results into • affected by • depends on • therefore • can stretch • causes

Table 5.17: List of causal based linking words.

Chapter 6

Studies Using Re-representation

6.1 Introduction

In this chapter, we address research questions 3 (feasibility, efficacy study) and 4 (proximity, expertise indicator study). Feasibility and efficacy studies are conducted to explore the ease of using the well-defined linking words by students. Proximity studies are conducted to explore the similarity and closeness of well-defined linking words with experts' usage in textbooks and ontologies. We wanted to explore the use of the re-representation method for its ease of use, efficacy with students and if it can bear any significance as an indicator of expertise. In this context, we discuss these studies with analysis and findings in this chapter.

6.2 Efficacy Study – 1

We wanted to explore if the re-represented concept mapping (RCM) can be used in the classroom and whether it is feasible and easy, given that it has constraints on the usage of linking words. This study aimed to test the RCM method for its ease and feasibility among high school students.

We conducted a study to determine if a constraint set of linking words is provided and if it would be difficult or easy for students to develop a consistent representation without any loss of expression. The study had a homogeneous sample of three groups of students (age 13-14 years, mixed gender) studying in Class 9 from a local urban school. The chapter on "The Fundamental Unit of Life" from Class 9 Science Textbook (NCERT, 2007c), was chosen for the study. All students were assigned the same task – to describe "the structure and function of nucleus and mitochondria". The three groups (in three different classrooms) were assigned three distinct modes of representation – description (DES), standard concept mapping (SCM), and re-represented concept mapping (RCM).

The research design comprised a one-shot study. Group 1 (n=32) was asked to complete the task using description (DES) mode using simple sentences (without any constraints). Group 2 (n=30) was assigned to achieve the same task using the SCM (without any constraints, seed concept terms provided, relation terms not provided). Group 3 (n=30) was assigned to complete the same task using the RCM (with constraints, seed concept terms, and a set of relation terms provided). The students were already taught the above chapter by their science teachers as per their classroom schedule. However, to help with recall, the same chapter was read out for all three groups. Before assigning the task of representing the domain for groups 2 and 3, an introduction, familiarization, and practice session of the concept mapping technique was conducted.

During the familiarization task, we explained the students about the meanings and distinctions of the linking words. We also explained their usage based on the role of the concepts. For example, to use 'consists of' to connect a part-whole and if the two concepts convey structure terms, to use 'has function' to connect a structure and process term, etc.

We used two entire class periods (of 40 minutes each) for the familiarization task wherein we explained the concept mapping technique to group 2. For group 3 also we used two classes for familiarization of concept mapping and linking words. Both the groups also did practice tasks during the familiarization.

This study used only a minimal set of relation names while constructing the maps for a given domain, as shown in Table 6.1. Given an SCM, it is possible to obtain an RCM by replacing the relation name, keeping the concept names more or less constant.

Dimensions	RLWs
Part-whole	consists of / part of; composed of
Class-inclusion	includes
Spatial inclusion	surrounded by; enveloped by; located in
Function	has function
Attributes	has nature; has size; has shape; has color; has property
Examples	example; instance of

Table 6.1: List of relation names provided with RCM for the domain.

Figure 6.1 shows a re-represented concept map on 'nucleus,' and Figure 6.2 shows a re-represented concept map on 'mitochondria,' both represented using a minimal set of linking words.

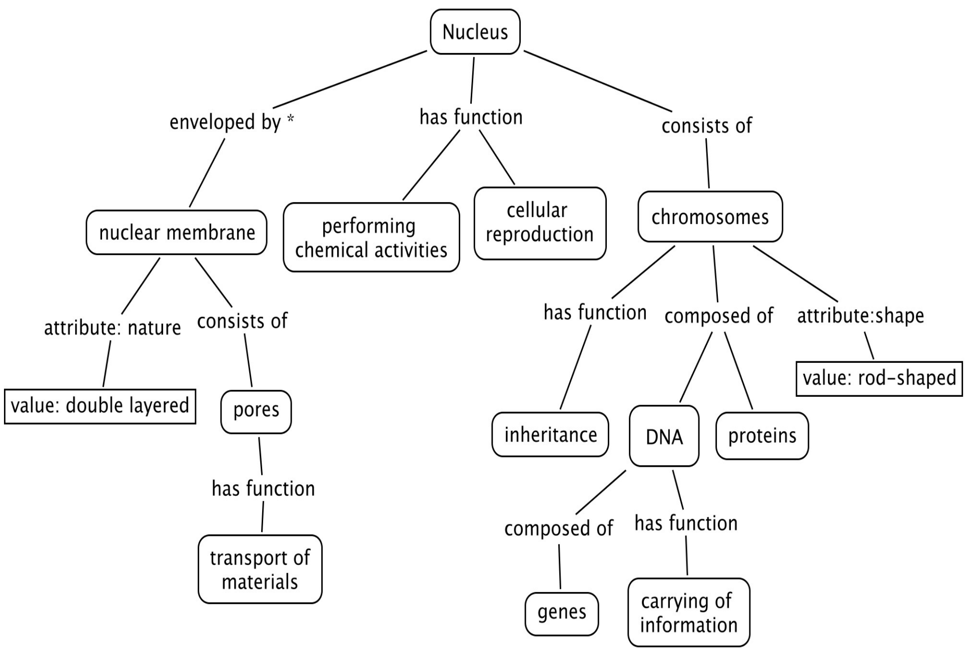


Figure 6.1: A re-represented concept map on 'nucleus' created using the minimal set of linking words.

Figure 6.3 shows students' representations of 'nucleus' and 'mitochondria' created using a minimal set of linking words from the group of RCM.

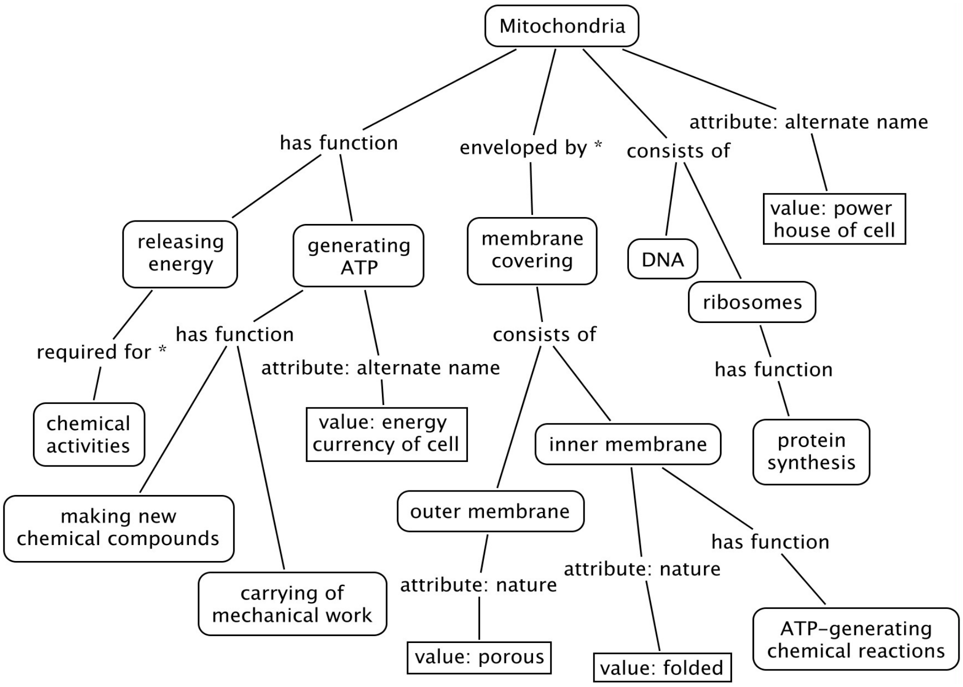


Figure 6.2: A re-represented concept map on 'mitochondria' created using the minimal set of linking words.

6.2.1 Analysis and Findings

The units of analysis were concept terms and relation terms in the domain. From the propositions, the concepts and relations were scored and analyzed. The chapter from the textbook was considered to be a control for group 1 (DES). A criterion standard concept map and a criterion re-represented concept map were used as controls for groups 2 (SCM) and 3 (RCM), respectively. The criterion concept maps shown in Figure 6.1 and Figure 6.2 were collaboratively created by researchers and the class 9 teachers. Almost all the critical concepts and propositions that were required to represent the domain were represented by all three groups.

The relation terms words were categorized into dimensions such as – part-whole, class-inclusion, spatial-inclusion, function, and attributes. A selected list of propositions from the three groups with a comparison with expert's text (Campbell and Reece, 2008) is shown in Table 6.2.

An account of the kinds of relation names used in all three groups was considered. During the analysis, the number of relation names (linking words) used for a given

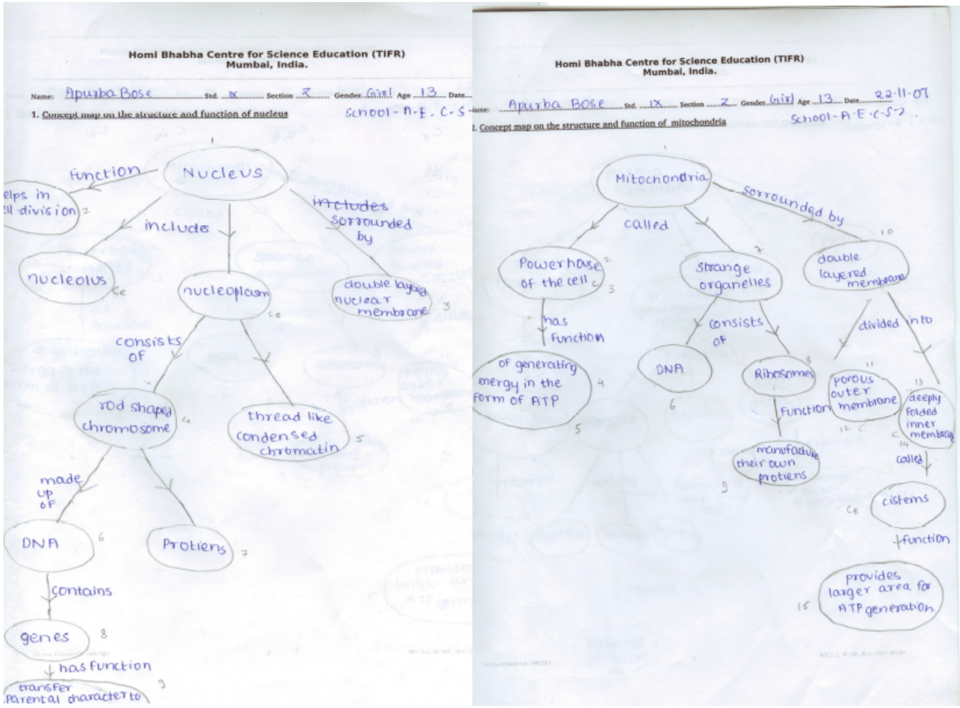


Figure 6.3: A class 9 student's RCM on 'nucleus' and 'mitochondria' created using the minimal set of linking words.

proposition by all three groups was compared. For instance, the sentence "nucleus is surrounded by nuclear membrane" was represented using different kinds of relation names such as – "made of", "consists of", "contains", "has", "divided into" groups 1 and 2.

The study revealed that the students from group 1 used 30, group 2 used 42, and group 3 used 7 (already provided) relation names to represent the same domain. As there was no constraint on relation names, the sentences were ambiguous, and therefore, there was no parsimony in the map as seen in groups 1 and 2. In these two groups, there appears to be less number of correct relations than those in group 3. Therefore, this indicates that the constraints did not hinder the expression of knowledge, as it was found that group 3 had more correct and valid relations even with a fixed number of relation names.

In the further analysis of the study, the number of incorrect relations was analyzed based on the incorrect choice of relation names. These were found to be more in groups 1 (average for nucleus = 1.73; average for mitochondria = 0.4) and 2 (average for nucleus = 1.16; average for mitochondria = 0.48) than those found in group 3 (average for nucleus = 0.66; average for mitochondria = 0.19). Concerning the depiction of incorrect relations,

Dimensions	Group 1 (DES)	Group 2 (SCM)	Group 3 (RCM)	Expert
Part-whole	nucleus is comprised of DNA; nucleus consists of DNA; nucleus contains DNA	nucleus has DNA; nucleus is made up of DNA	nucleus consists of DNA	nuclei contain DNA
	mitochondria has its own DNA and ribosomes; mitochondria consists of its own DNA and ribosomes; mitochondria contains its own DNA and ribosomes	mitochondria have DNA and ribosomes ;mitochondria made up of DNA and ribosomes; mitochondria consists of DNA and ribosomes; mitochondria contains DNA and ribosomes	mitochondria consists of DNA and ribosomes;	mitochondria contain DNA and ribosomes
Spatial-inclusion	mitochondria is a double layered organelle; mitochondria is a cell organelle which is double layered membrane; mitochondria has double membrane; mitochondria divided into double membrane	mitochondria have two membrane; mitochondria consists of two membrane, outer and inner the double layered membrane is made up of outer and inner membrane mitochondria has double layered covering	mitochondria is surrounded by double membrane	mitochondria is surrounded by double layered membrane

Table 6.2: A comparison of few sentences over the different groups with expert's representation.

it was observed that the more the number of linking words used, the more it was prone towards inaccuracy. Group 3 used a small set of relation names, and it had a lesser number of incorrect relations as compared to groups 1 and 2.

We also highlight another observation that: in the description mode, there were quite a few misconceptions and idiosyncratic ideas seen in group 1. Some of these were — "if the nucleus is removed, the protoplasm dries up, and the cell dies"; "mitochondria has genetic material like ribosomes"; "if mitochondria are removed from the cell, it will not get energy and will dry and die"; "nucleus contains membrane-bound structure called chromatin"; "nucleus contains ribosomes". Some from group 2 used long sentences into concept names and relation names in a linear way. Whereas in group 3, there were almost no such misconceptions or idiosyncratic representations. The RCM method imposes a constraint which, in turn, facilitates students to represent knowledge more meaningfully, without any loss of knowledge.

For all three groups, a score of 1 each was assigned to each non-redundant concept and a score of 1 each for the valid relation of nucleus and mitochondria. The scores are indicators of students' understanding of the domain. We believe that the scores for concepts and relations are proportional to their understanding. The students' scores were compared with criterion or expert based concept maps on nucleus and mitochondria as shown in Figure 6.1 and Figure 6.2. These maps were created collaboratively with teachers and researchers.

According to the concept mapping scoring method, concept connecting with linking words to other concepts form a meaningful proposition or a statement. This is considered as the smallest unit of knowledge. Assigning points for concepts is about the technical knowledge about the topic, and assigning points for linking words is only when it forms a meaningful statement. So although the score is for linking words but it is assigned based on the validity of the whole statement and all those statements in the concept map. Therefore we suggest that it corresponds to their understanding of the topic. Because our work focuses on the linking words, we give special attention to the kinds and nature of linking words used to form those statements.

The data were treated for parametric tests. A single-factor analysis of variance

(ANOVA) was used to compare the variance of the three groups. For the node names of nucleus the $F(2, 89, 91) = 2.66$, $p>0.05$ was found non-significant. However for relations of nucleus the $F(2, 89, 91) = 8.20$, $p<0.05$; and for the node names of mitochondria the $F(2, 89, 91) = 4.13$, $p<0.05$; and for the relations of mitochondria the $F(2, 89, 91) = 4.50$, $p<0.05$ were found to be significant (Figure 6.4).

Domain	***Mode***	***N***	***Mean***	***SD***	***Variance***	***SS***	***DF***	***MS***	***F***
Nucleus --- Concepts	DES	32	10.88	2.85	8.11	34.29	2	17.14	2.66
	TCM	30	9.43	2.16	4.67	574.37	89	6.45	
	RCM	30	10.5	2.54	6.47	608.65	91		
Nucleus --- Relations	DES	32	6.78	2.01	4.05	97.5	2	48.75	8.20 *
	TCM	30	7.37	2.11	4.45	529.24	89	5.95	
	RCM	30	9.2	3.08	9.48	626.74	91		
Mitochondria --- Concepts	DES	32	11.31	2.81	7.9	70.85	2	35.42	4.13 *
	TCM	30	9.23	3.18	10.12	763.71	89	8.58	
	RCM	30	9.87	2.79	7.77	834.55	91		
Mitochondria --- Relations	DES	32	6.09	1.57	2.47	50.71	2	25.36	4.50 *
	TCM	30	7.07	3.06	9.37	501.29	89	5.63	
	RCM	30	7.9	2.29	5.27	552	91		

** $p < 0.05$; F Critical = 3.10*

Figure 6.4: ANOVA test.

To further analyze which of the three groups produced significant results a t-test was performed. We have found significant differences in the relations that have been depicted for nucleus $t(49.45) = 3.6$, $p<0.05$, and mitochondria $t(50.93) = 3.59$, $p<0.05$. In these two cases, the refined concept mapping is significant over the description mode and traditional concept mapping mode $t(51.30) = 2.6$, $p<0.05$. As far as the depiction of node names, there have been no significant differences in the traditional concept mapping method $t(56.52) = 1.75$, $p>0.05$ and the refined concept mapping method $t(57.02) = 0.8$, $p>0.05$, which shows that the refined concept mapping does not affect the representation of critical concepts (Figure 6.5). The mean of the DES group for Mitochondria (Concepts) is higher than the other two groups. This is because we analyzed the scores for concepts also in addition to relations. Although our focus is on the scientifically accurate relations (in terms of statements), we also wanted to demonstrate that the critical concepts that are required for the topic of Mitochondria are not missed out or there is no loss of knowledge of concepts. Through this we explain that given the concepts that are required to represent the topic the relations can be depicted by connecting those.

Domain	Mode	N	Mean	S.D.	t
Nucleus --- Concepts	DES	32	10.88	2.85	2.25 *
	TCM	30	9.43	2.16	
	TCM	30	9.43	2.16	1.75
	RCM	30	10.5	2.54	
	DES	32	10.88	2.85	0.54
	RCM	30	10.5	2.54	
Nucleus --- Relations	DES	32	6.78	2.01	1.11
	TCM	30	7.37	2.11	
	TCM	30	7.37	2.11	2.6 *
	RCM	30	9.2	3.08	
	DES	32	6.78	2.01	3.6 *
	RCM	30	9.2	3.08	
Mitochondria --- Concepts	DES	32	11.31	2.81	2.72 *
	TCM	30	9.23	3.18	
	TCM	30	9.23	3.18	0.8
	RCM	30	7.77	2.79	
	DES	32	11.31	2.81	2.03 *
	RCM	30	7.77	2.79	
Mitochondria --- Relations	DES	32	6.09	1.57	1.55
	TCM	30	7.07	3.06	
	TCM	30	7.07	3.06	1.19
	RCM	30	7.9	2.29	
	DES	32	6.09	1.57	3.59 *
	RCM	30	7.9	2.29	

** $p < 0.05$; t Critical = 2.00*

Figure 6.5: T-test.

We think that ambiguity was reduced in group of RCM because we suggested students in this group to use and re-use the provided linking words for same meaning. Interestingly, it was observed that there was no difficulty for group 3 (RCM) to learn and use the constraint-based concept mapping during the study (Kharatmal and Nagarjuna, 2009). In fact the students felt at ease and were happy when they were provided with cues for concept and relation names. This reminds us of Ausubel et al. (1978) classroom theory about anchoring a new concept with the already existing concepts. From the study, it can be observed that RCM is parsimonious, and it does not hinder the representation of critical concepts. This has already helped in achieving the primary objective of the study.

6.3 Efficacy Study – 2

Considering the significance of focusing on linking words, we investigated the effect of providing linking words in concept mapping for representing cell biology concepts. We conducted a preliminary study with nine college biology students (18-20 years of age). We conducted a familiarization task of introducing the concept mapping method and its elements of concepts, propositions, branches, hierarchy, cross links, and scoring rubric. Students also worked out certain trial or practice concept maps on topics of 'food', 'transport system', from which it was ensured that students have learned the concept mapping tool. Further, a textbook (NCERT, 2007c) passage on plastids was used to create concept map with the traditional method without providing linking words. This was followed by creating concept maps by providing linking words: part of/consists of, includes/kind of, contains, known as; has function/has role; has shape; has size; has color. A criterion concept map on plastids was created by the teachers and researchers collaboratively with consensus (Figure 6.6). It can also be called as an expert's concept map or a standard concept map.

Students' concept maps were evaluated and scored based on the established scoring rubric (Novak, 2010; Novak and Gowin, 1984) and compared with the scores of criterion concept map. The concept maps were scored for structural complexity (the total number of propositions used) and propositional validity (number of valid propositions/total number of propositions) (Briggs et al., 2016; Mintzes et al., 2001). The propositions were scored based on the usage of provided linking words used to connect concepts. For example, the usage of linking words "contains" would be considered valid in connecting "plastids" and "chlorophyll pigments" and invalid in connecting "plastids" and "ATP. Therefore a valid proposition is considered as "plastids contains chlorophyll pigment" whereas an invalid proposition would be "plastids contains ATP". All the valid propositions in the maps were scored for 1 point.

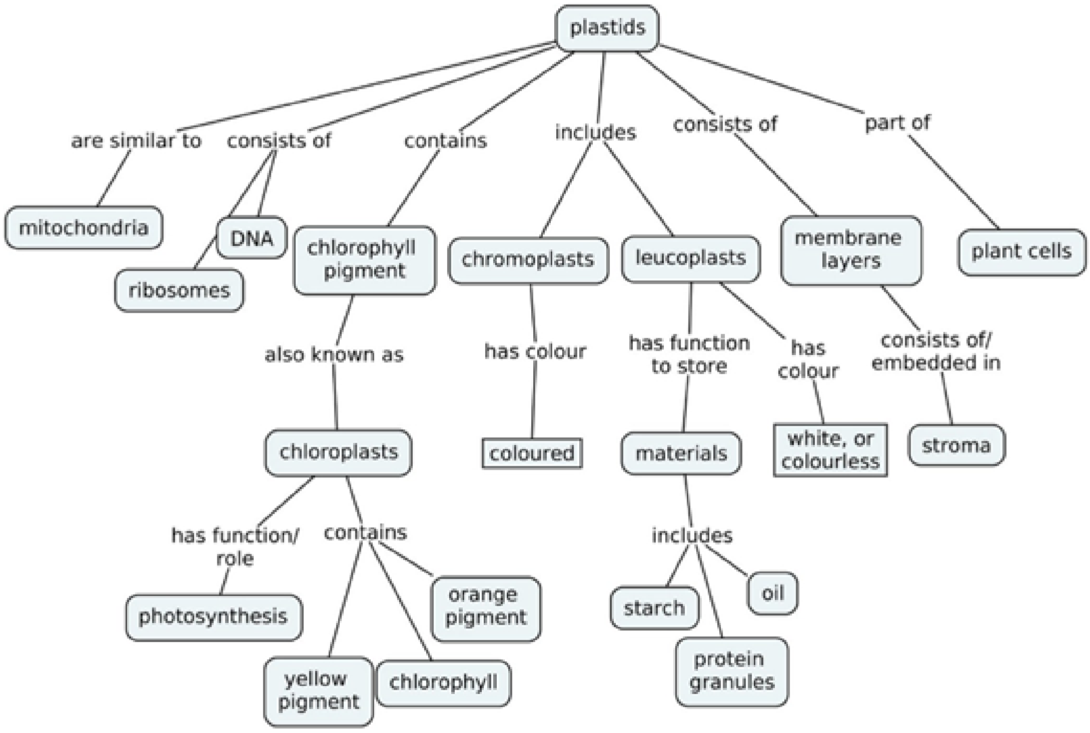


Figure 6.6: A criterion concept map on the topic of Plastids created in collaboration with teachers and researchers.

Figure 6.7 shows a representative student's concept map on plastids without provided linking words. While Figure 6.8 shows a representative student's concept map on plastids with provided linking words.

6.3.1 Analysis and Findings

The list of 20 propositions from criterion concept map on plastids was used to compare the valid propositions of each students concept maps from both the methods of without providing and with providing linking words. The score for propositional validity was used to determine a proximity percentage to criterion concept map by counting the percentage from total number of valid propositions that were matching with the criterion concept map (20 number of propositions). For example, if a student has a total proposition of 13 and from these there are 9 invalid and 4 valid propositions (as in criterion concept map), then the proximity percentage was calculated as 31% (Kharatmal and Nagarjuna, 2024).

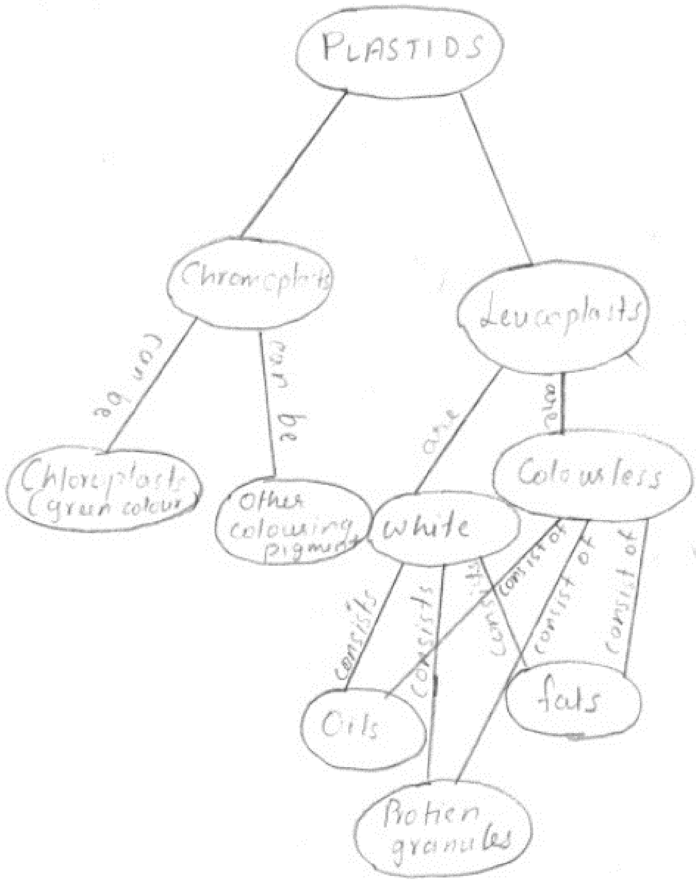


Figure 6.7: A representative student's concept map on plastids (no linking words provided).

The Figure 6.9 showed an increased percentage due to the usage of the provided linking words that enables to create more valid propositions. The findings from this study indicated concept mapping with provided linking words facilitated students' representations depicting closer to expert's (criterion) representations (Kharatmal and Nagarjuna, 2024).

6.4 Proximity Study – 1

A measure of proximity can be in terms of: a text requiring lesser re-representation to be considered as more proximate, whereas the text requiring more re-representation can be considered as least proximate. This can yield a *proximity percentage* of the text. We address the proximity study by analyzing if the choice of predicates can characterize expertise. We consider the re-represented method's re-represented linking words (RLWs) as an indicator of expertise

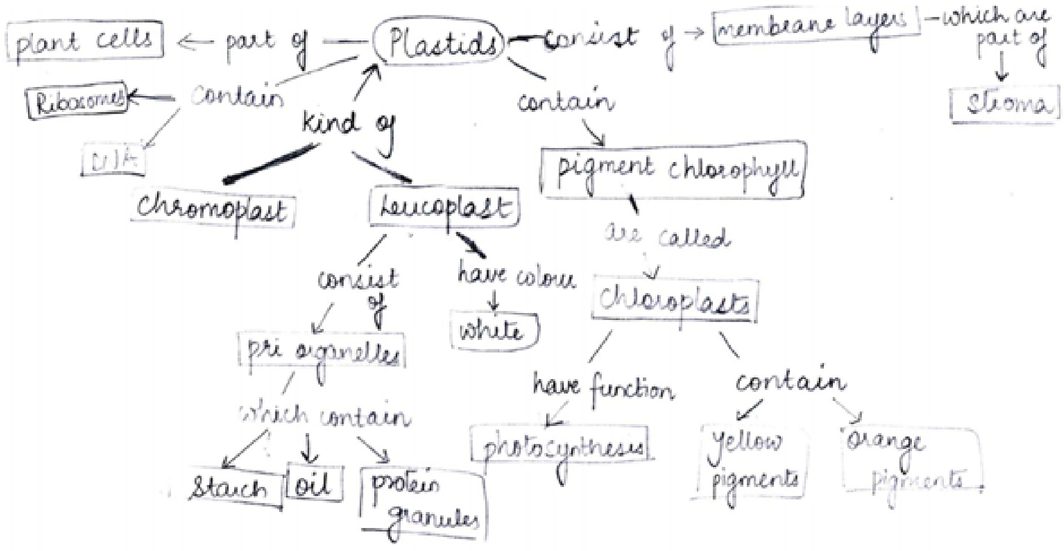


Figure 6.8: A representative student's concept map on plastids (with linking words provided).

The topics on 'nucleus' and 'mitochondria' were analyzed from three books UG-1 (Taylor et al., 2003), UG-2 (Campbell and Reece, 2008), UG-3 (De Robertis and De Robertis, 1985) as shown in Table 6.3.

We also compared the school text with the expert's text for the choice and usage of predicates. The most widely and internationally used textbook for undergraduate biology is Campbell's Biology (Campbell and Reece, 2008). We compared the three school and high school levels textbooks of classes 8, 9, and 11 for their usage of verbatim choice of predicates as well as re-represented predicates. We compared those with Campbell's text (expert's) choice of verbatim (original) predicates. The whole count of predicates in each dimension was studied to know how many from this count would be proximate with Campbell's verbatim use of predicates.

6.4.1 Analysis and Findings

The proximity index was calculated as the number of common predicates between the verbatim predicates of texts and re-represented predicates, divided by the number of verbatim predicates for a topic. Based on this, the proximity (or overlap) of the verbatim predicates with the Reference Set showed an increase from UG-1 to UG-2 to PG (UG-3).

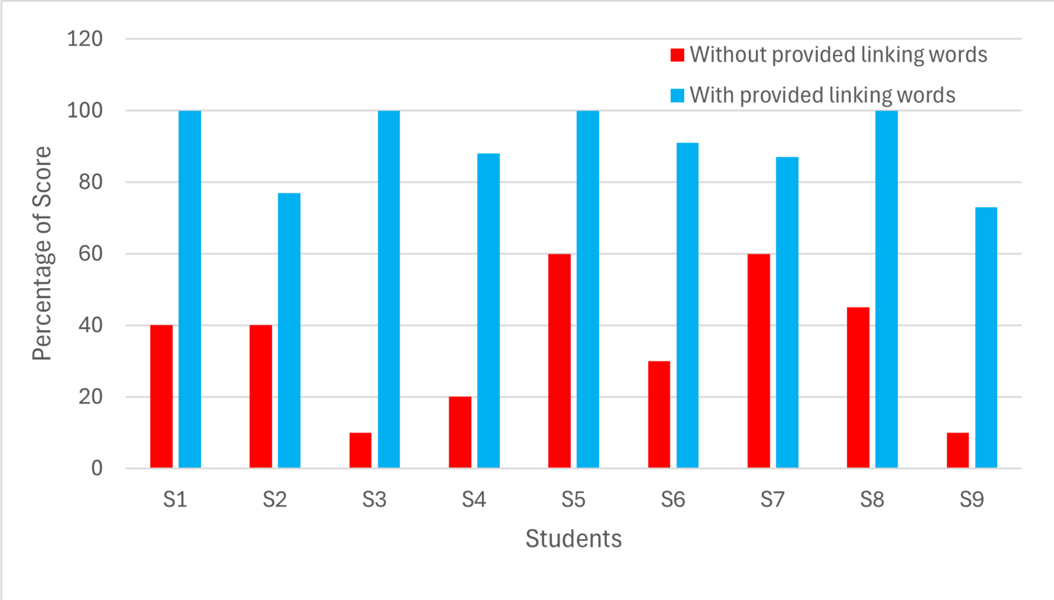


Figure 6.9: A graph showing increased percentage of propositions with provided linking words when compared with criterion concept map.

The UG-1 level text already showed a higher proximity of 83 % (Figure 6.10). However, it was not the case with school texts, where the proximity was found to be just about 50 % (see Chapter 3). This study demonstrates that the proximity of predicates of texts and Reference Set correlates with expertise. This implies that by applying the re-represented relations, the text can become closer to an expert representation and can bear an indicator of expertise. This study is published in Kharatmal and Nagarjuna (2016). In a similar way, students' representations can be compared with expert's representations to study a degree of novice-expert transformation.

We consider Campbell's Biology (Campbell and Reece, 2008) as an expert-level textbook and compared the school and high school level text's RLWs with it. We listed the verbatim usage of LWs in Campbell text (SCM) and compared those with the LWs and RLWs of the school and high school levels i.e., classes 8, 9, 11 texts. The linking words in each dimension were counted and compared with the whole count of LWs within the same dimension. This count is compared to find the proximity with Campbell's verbatim text and usage of linking words. The graph in Figure 6.11 suggests that RLW has increased proximity with that of the experts' text (verbatim) in the usage of linking words.

When the verbatim linking words (LWs) are compared with experts' verbatim linking

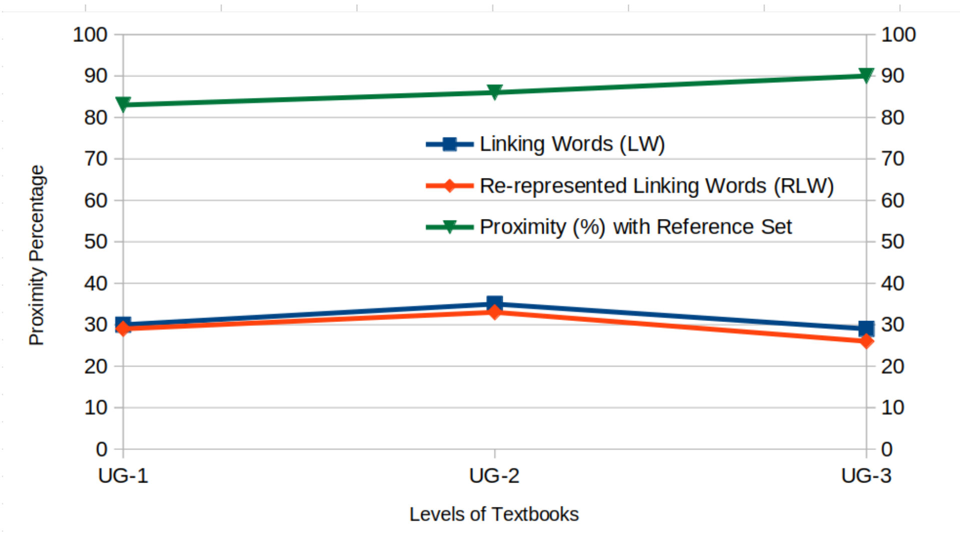


Figure 6.10: Graph depicting the higher degree of proximity percentage. The secondary Y-axis depicts the number of LWs and RLWs. (Note: The lines connecting points are just a guide to the eye).

words (LWs), and when the re-represented linking words (RLWs) are compared with experts' verbatim linking words, we can see that the increasing order increases by a range of at least 10-20 %. There is a higher degree of proximity as compared to the verbatim linking words to re-represented linking words. This shows that certainly if and when we do re-represented, it becomes much closer to an expert's representation. This shows that characterizing linking words can become an indicator of expertise (Kharatmal and Nagarjuna, 2008).

6.5 Proximity Study – 2

The ontology has domain-specific content classified into class terms, object property, annotation (Ong et al., 2017). The class terms and object properties are well-defined by the community of subject experts and knowledge engineers in the ontology database. We compared the re-represented linking words with the ontology repository's object and data properties.

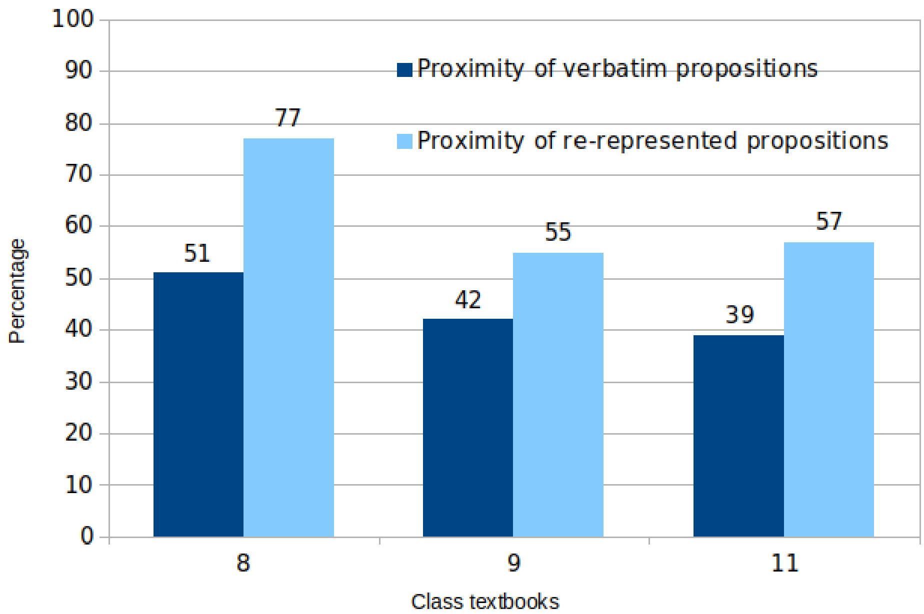


Figure 6.11: Increased proximity of re-represented linking words with Campbell's usage of verbatim linking words.

6.5.1 Analysis and Findings

In this method, we compared the re-represented predicates with the expert's repository of the published relations and attributes (Ong et al., 2017). The verbatim and re-represented predicates for five levels of text were compared from the published repository of relations (object property) and attributes (data property). The comparison depicted increased proximity as shown in Figure 6.12.

One important aspect in doing this comparison is validating the re-represented linking words from the ontology. The classification of terms in ontology complements the scheme of – concepts, relations, and attributes – followed in our content analysis. We sought to compare the LW and RLW to the class term, object property, and annotation property from various ontologies. Each of the LW and RLW, for all five levels of text, were compared from the ontology as shown in Figure 6.12.

In the attribution category, we observed that the RLW attributes 100 % match with the ontology database.

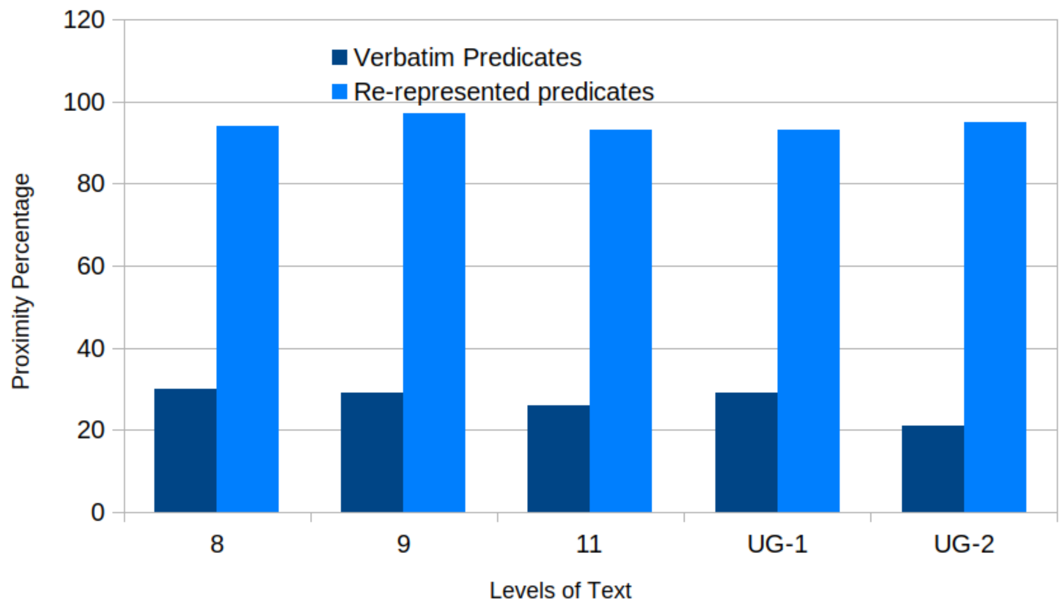


Figure 6.12: Higher proximity percentage with expert's published repository.

6.6 Summary

In this chapter, we presented two studies - feasibility and proximity – to explore possible uses of the re-representation method. Needless to say, both these studies need to be replicated with more extensive empirical work, which we wish to carry on in future. These studies have a limited role of indicating possible ways in which the method could be used and to test the conjectures that emerged during the study. We are fully aware that they fall short of validating the methodology developed.

In the next chapter, we discuss the conclusions and implications of the thesis.

Dimension	Linking Words UG-1	Linking Words UG-2	Linking Words UG-3
Part-whole	composed of (1), exists (1), in form of (1), may be present in (1), not consists of (1), perforated by (1)	consists of (9), is perforated by (1), made of (5)	composed of (1), consists of (5), is associated with (1), is interrupted by (1), present/seen in (3)
Class-inclusion	includes (1), kind of (1), type of (2)	divides into (1), includes (1), kind of (2), type of (1)	divides into (1), includes (1), type of (1)
Spatial-inclusion	bound to (1), contains (11), covered with (1), is continuous with (1), located inside (1), occur near (1), packed in (1), surrounded by (2), wound around (1),	are continuous (1), contain (6), enclosed by (5), exit through (1), extends through (1), is lined by (2), is organized (1), separated by (2), located (3), occurs in (1),	aligned (1), attached to (4), contains (10), continuous (1), encloses (1), located between / are situated between (2), occupied (1), packed in (1), projects into (1), separated by (2), surrounded by (3), is traversed by (1)
Function	are sites of (3), formed by (1), has function (6)	are sites for (2), has function (11), play role (1), synthesized by (1),	is site for (1)
Attribution	has diameter (1), has form (1), has length (1), has property (1), has shape (1), has size (1)	appears as (1), has form (3), has length (1), has nature (3), has number (2), has property (2), has size (1)	has form (1), has nature (3), has shape (1), has size (1)
Others	called (5), means (1)	called (1)	called (1)

Table 6.3: List of linking words used by college textbooks at UG-1, UG-2, UG-3 levels for 'Mitochondria' and 'Nucleus' topics. The frequencies of each linking word are indicated in

Chapter 7

Conclusions and Implications

7.1 Introduction

The thesis had two objectives – (i) Designing and developing of the Re-representation method; (ii) Using the Re-representation method for semantic content analysis or relational content analysis of science text. The thesis began with motivation and identifying gaps that led to the framing of four research questions with the first two research questions complementing the two objectives of the thesis (RQ1 in Chapter 3 and RQ2 in Chapter 4 and Chapter 5. The latter two research questions have been addressed as part of two studies on feasibility, efficacy (RQ3 in Chapter 6) and proximity (RQ4 in Chapter 6).

A graphical representation of the research model of the thesis is presented in Figure 1 in the executive summary.

7.2 Conclusions from the Thesis

We draw conclusions from this exploratory work conducted as part of thesis by recalling the four research questions and answering them on the basis of the findings presented in the previous chapters.

RQ1: Can we design a method for representation of biology language focusing on the semantics of predicates? What features of biology language could emerge from it?

The Re-representation method has been designed and developed in the form of iterative refinements from Phase 1 to Phase 6, though not exhaustive. These phases were not linear in development and in implementation. The theoretical frameworks of knowledge as a network and Representational Redescription; and of biomedical ontology, nominalization, and conceptual spaces are complementing for developing the phases of the re-representation method focusing on semantics. We addressed the nature of biology language in terms of structure-function, variable properties, processes for scientific explanation, and procedures for experimentation.

RQ2: Could this method be used for the semantic content analysis of biology language? What would the analysis unravel about the nature of biology knowledge?

The re-representation method was used for content analysis of 6 texts from school, college, undergraduate, and postgraduate levels and high school and undergraduate students' descriptions, focusing on the semantics. This exploratory study has resulted into significant outcomes, findings and patterns.

Outcomes:

The content analysis has led to outcomes that provide insights into the characterization of biology knowledge. A major outcome of the study has been the emergence of four ontological categories – *structure based; class based; process based; attribution based* for representation of science text. The study has also helped to emerge distinct categories of linking words within these four ontological categories. These are:

- *structure based – part-whole, material-composition*
- *class based – class-subclass*

- *process based – process, change and development, movement, causal, function*

- *attribution based – measurable, form, physical quantity*

Patterns:

The study of using the re-representation method in content analysis has led to the emergence of certain significant patterns as we addressed the research questions. The patterns indicate that:

- Even though the concepts increase across various levels of knowledge, the linking words attain saturation. As part of the content analysis of the texts, there is prominence of function based linking words and form category of attribution. Interestingly, as the levels of texts change from school to college, there is an increase in the process based linking words.

- When the verbatim linking words (LW) and the re-represented linking words (RLW) are compared, there occurs a reduced number of RLWs. Not only the RLW is reduced, but the ratio between the LW and RLW achieves saturation. Due to the re-representation of knowledge, it leads to parsimony during mapping knowledge.

- An interesting pattern from the content analysis is about the nominalization of process-based linking words. This leads not only towards attaining saturation of function-based linking words but also towards disambiguation and differentiation of process based linking words. This characteristic is significant as it would help towards a process-centric approach to biology knowledge.

Characterization:

During the re-representation of knowledge, the nature of changes in the linking words occurs by – *category shift, coalescence, differentiation, disambiguation, nominalization.* This nature of change can suggest a model for the restructuring of knowledge.

RQ3: Could this method be feasible? To what extent using the finite set would be effective in representing biology knowledge?

The method of using a constraint set of linking words can facilitate representing valid and accurate propositions.

RQ4: In what ways would the semantic content analysis serve as a measure of expertise? Could this method be used as proximity indicator?

The re-represented linking words (RLWs) can be considered as an indicator of expertise. The significance of using RLWs during the re-representation of text is that the RLWs were found to show proximity not only to the expert's text but also to the expert's representation of ontology. This would also signify that if RLWs are used during mapping, it can achieve expert-like representation.

7.3 Contributions of the Thesis

We discuss the applications and extension of this contribution to other areas.

The Re-representation Method

The design and development of the re-represented method focusing on semantics is a novel contribution of the thesis. The method has features for representing various forms of cell biology knowledge and thus characterizing cell biology knowledge.

A Methodology for Content Analysis

A methodology focusing on semantics in a scientific language for content analysis is a contribution from the thesis. This methodology can be applied to textbook analysis, conversation analysis, students' descriptions, audio-video transcription, research articles, scientists' talks, etc., depending on the objective.

A Reference Set, Semantic Categories and Re-representation Rules

A Reference Set of well-defined relations and ontological categories (see Appendix) and semantic categories and re-representation rules (see chapter 3, section 3.9) are contributions of this thesis.

Mapping Biology Knowledge

The characterization of biology knowledge has contribution towards emerging categories and sub-categories that can be applied during mapping. These categories can be considered necessary and sufficient to represent any biology text with topics related to cell structure and function. The four ontological categories and their distinct sub-categories are not to be considered as discrete but as complementary to the existing categories.

Concept Mapping

The contribution of the study bears significance for the discipline of concept mapping. Just as seed concepts are provided in standard concept mapping, seed relations can also be provided during re-represented concept mapping, which can be applied as additional characteristics to focus on the linking words during mapping. The study has helped to provide additional rubrics/criteria for focusing on linking words during mapping knowledge even with the standard concept mapping method. The hierarchical organization of knowledge is another distinguishing characteristic of expert knowledge. While talking about the need for hierarchy in concept maps, Novak and Gowin (1984) refers to (Ausubel et al., 1978) notion of subsumption. In a concept map, relations used in a hierarchy are not necessarily logically transitive. These additional conventions facilitate identifying hierarchical relations that are logically transitive since the same linking word is used for a given intended meaning. This affords a new scoring measure for counting hierarchy.

The Re-representation method's phase 2 and phase 3 are contributions towards representing variable properties and dynamic propositions. This can pave the way toward a process-centric biology while creating concept maps. Further, the nested node feature of the CmapTool can be utilized for representing processes as state-change.

7.4 Discussion

Dismabiguation

In all human languages, even in different cultures, certain words can mean different in different contexts. Such words are potentially 'ambiguous'. In this regard, word sense disambiguation (WSD) is about deciding which specific meaning is intended out of several meanings of a term in a given context. Disambiguation occurs from various sources – syntax, word meaning, knowledge, and context. Here the context is the neighboring words, relations. Therefore, researchers apply the semantics, or semantic relatedness in a network, use the knowledge base of thesauri, WordNet for WSD (Roget, 1962; Fellbaum, 1998; Miller, 1995; Yang, 2006; Tsatsaronis et al., 2008). Our approach is also influenced by this research on semantic relatedness for context, a network of neighboring concepts for providing meaning in science text.

We acknowledge that ambiguities may be perceived differently by different audiences in different cultures. Sometimes, with ambiguities, there is scope for creativity, and poetry that are open for interpretations differently. Sometimes, the ambiguities when placed within a context, help in precision in conversation. However, when the domain is about science, the principles of disambiguation, consistency, parsimony, and rigor become the hallmark of scientific language. With this criteria, we place the disambiguation principle as important for re-representing text. Its most efficient application is in scientific communication, and air traffic communication, where disambiguity and precision are of utmost importance for human survival.

Conceptual Modeling

Conceptual modeling deals with creating representations for understanding and communication (Guizzardi and Halpin, 2008), and our work addresses this for scientific communication i.e. language of science in text. It is required that ontologies capture formal systems defined by concepts and relations, and specifications need to be precise, unambiguous, and consistent (Hofmann, 2013, pp. 81-82). In conceptual modeling, knowledge representation, specifications of core semantic relations such as part-whole relations are being presented with the formal specifications to be applied for reasoning and inferencing (Keet and Artale, 2008). Just as the ambiguity of part-whole relations is documented in the ontology modeling area, similar problems exist in the language of science used in textbooks. In this work, we have not only presented some of these problems but also have created re-representations of text, following the applications of such formal specifications.

Nuclear Predicate Vocabulary

Generating a list of nuclear predicates for each domain is possible, and it can be realized from the published ontologies. In the semantic spectrum, the number of predicate terms used in formal expressions is fewer than those in lesser formal expressions (McGuinness, 2003). In the Controlled Natural Language (CNL) area of work, researchers have used a constrained set of well-defined predicates in the ontology for creating unambiguous relations (Clark et al., 2010). Knowingly or unknowingly the value of minimalism of predicate terms is promoted in rigorous subjects. Such a finite set can facilitate for creation of a vocabulary of nuclear predicates just as Stubbs (1986) argued for creating nuclear vocabulary. Given the possibility of published ontologies contributing to nuclear predicates for a given domain, we can rewrite scientific text using only the set of nuclear predicates.

Growth of Knowledge

As illustrated in the thesis, we eliminated the ambiguity and made the meaning explicit by re-writing. By doing this, we demonstrated that meaning comes from predicates while class terms derive meaning from them. This is of significance in the growth of knowledge because for any new knowledge the same set of nuclear predicates can be reused for meaning. The saturation of predicates bears value for disambiguation using "one predicate for one meaning" throughout, for consistency using "same predicate for same meaning". Within the relations, linking words (object properties) attained saturation, but we also found that the attributes (data properties) showed an increase as the levels of text increased (Kharatmal and Nagarjuna, 2011). As we advance into college-level text, the depth and breadth of the subject increase, which are indicators of the growth of knowledge. Using ontological descriptors, we can say that the depth increases due to the parameterization of the phenomena, while breadth increases due to the introduction of new class terms. Considering the economy of expression as a virtue sought after in scientific modeling, attributes (data properties) also tend to be not very high in number. We suggest that making the distinction between relations and attributes (i.e. object property and data property) would make the data property (label) explicit. Most importantly, this method is applicable for process modeling. This finding can bear significance for model-based reasoning, depicting the change in process representation suggesting that the core of scientific knowledge lies in the modeling of a domain (Kharatmal and Nagarjuna, 2013).

Automation of Re-representation Rules

From the content analysis method, the coding pattern can be used for machine learning, automated analysis of text, etc. During the analysis, the coding emerged a consistent pattern. Given any sentence that uses the verb "has" for the intended sense of a measurable data property (for example in (5) and (5′)), then the coding is *[R3, C8]*. This may be of significance for machine learning, to use the pattern for specific coding, such that whenever the coding *[R3, C8]* appears, the verb in the sentences depicting a measurable property is required to be re-represented with the appropriate data property. Similarly, as in (6), (7) and (6′), (7′), the verb "has" is used for the intended meaning of meronymic inclusion, then the coding pattern is *[R3, C1]*. Our method of relational content analysis comprises of re-representation rules and coding rules for disambiguation.

It can be applied for developing coding algorithms, for example in Automated Relation Extractor (Chaudhri et al., 2021). Our work can be considered parallel with the efforts of at least two forms of Controlled Natural Languages (CNLs): in Attempto Controlled English (ACE) (Fuchs et al., 2008; Schwitter, 2010; Kuhn, 2010, 2014), and Computer

Processable Language (CPL) (Clark et al., 2010). Although ACE is applied to the English language, our proposed method of re-represented is being developed specifically for scientific language in the text. Further, in the area of creating automated reasoning, relevant concepts and relations are considered for converting biomedical queries into Answer Set Programming (ASP) (Erdem and Yeniterzi, 2009). We propose that the set of relations extracted from the content analysis method can bear application for the automated reasoning algorithm.

If our argument based on the network theory of meaning is valid, it is essential to compile a dictionary of relations and attributes (see Appendix), so that students and teachers address these explicitly. Understanding their use helps them to construct meaningful assertions. This prescription can become almost mandatory in teaching explanatory models in science, as these are relations between attributes. We ask: Can these, then, become explicit objectives of education?

7.5 Reliability

The criteria of reliability in quantitative methodologies differ from those in qualitative methodologies, and therefore the canons of reliability for quantitative research may not work for qualitative research as suggested by LeCompte and Preissle (1993, p. 332). The difference in purposes of evaluating the quality of studies in quantitative and qualtitative research is also one of the reasons that the concept of reliability is irrelevant and sometimes misleading in qualitative research (Golafshani, 2003; Stenbacka, 2001). Researchers argued that as reliability concerns measurements, it has no relevance in qualitative research (Stenbacka, 2001), further stating that it is irrelevant in the judgment of the quality of qualitative research.

Reliability in qualitative research is not the same as for quantitative research (Cohen et al., 2015). Typically in quantitative methodologies, there occurs a degree of control and manipulation of phenomena, which can distort the natural occurrence of phenomena or patterns emerging in qualitative studies. The uniqueness of naturalistic studies is its strength (Cohen et al., 2015). Although the term 'reliability' is a concept used for testing or evaluating in quantitative studies, in qualitative studies the idea of testing is done by information elicitation, with the purpose of 'generating understanding' (Stenbacka, 2001, p. 551). In this context, the most important test of any qualitative study is its quality (Golafshani, 2003). A provocative statement in the defense of qualitative research states that "if a qualitative study is discussed with reliability as a criterion, the consequence is rather that the study is no good" (Stenbacka, 2001, p. 552).

There seems no consensus among researchers on how to best judge the quality of research based on content analysis (Bengtsson, 2016). The researchers are divided into two groups: one is those who debate the use of the same criteria and concepts as in quantitative research (validity, reliability and generalizability) (Downe-Wamboldt, 1992; Long and Johnson, 2000) – and the other group belong to those who believe that a different set of criteria and concepts is needed (Catanzaro, 1988; Graneheim and Lundman, 2004). In this context, Lincoln and Guba (1985) replaces the term 'reliability' and advocates the use of concepts such as "credibility", "dependability", "transferability" and "confirmability", in qualitative research. In a similar context, the notion of "dependability" is endorsed with the notion of "consistency" or "stability" in qualitative research (Seale, 1999). Further, Denzin and Lincoln (1998) suggests reliability to be replaced by replicability in qualitative research, which can be addressed by stability, when the researcher would have made the same observations and interpretation over a period of time. Conforming to this position, we re-iterate that as the Re-representation method was being developed over phases and also across different levels of textbooks over a period of long time, establishes replicability in our work.

Following the second group of school of thought, we prefer to apply the criteria of stability in our study. Stability is a tendency to consistently code and re-code the same data in the same way over a period of time. In our methodology, there is scope for coding and categorizing the same data in the same way over a period of time, because it follows the standard classification scheme (illustrated in detail in chapter 4 and chapter 5), by applying codes and re-representation rules, thus ensuring stability. A linking word will be categorized for meronymy repeatedly over a period of time because it follows the standard semantic relationship between a part and a whole. The researchers coded and categorized the data repeatedly over a period of time, during the phase-wise development of the re-representation method and were able to achieve the same categorization.

In content analysis, reproducibility is considered an important interpretation of reliability (Krippendorff, 2004). The methodology of content analysis can provide reproducibility because the general methods are applied to establish the categories. For example, the linking words and attributes are classified into various dimensions such as – meronymy inclusion, class inclusion, spatial inclusion, form, physical quantity, etc. (see chapter 4). The classification of linking words when assigning coding corresponds to a standard scheme, is said to be considered to achieve accuracy.

The process of content analysis (chapter 4) as explained in the sections from coding to assigning three levels of categorization, was carried out by two researchers. Wherever there were discrepancies in categorizations, those were resolved through consensus. The interpretations were made by the researcher and are reported in chapter 5.

7.6 Validity

The term validity in qualitative research is different from that in quantitative studies. Terms such as "credibility" and "authenticity" are more common in qualitative studies. Researchers have developed their own concepts of validity by applying quality, rigor, consistency, and trustworthiness (Davies and Dodd, 2002; Golafshani, 2003; Mishler, 2000; Seale, 1999; Stenbacka, 2001). In qualitative study, reliability is seen as a consequence of the validity in a study (Patton, 2002a,b). It is considered that when a study demonstrates validity, it can suffice to establish reliability (Lincoln and Guba, 1985, p. 316).

In qualitative research studies, the term validity is argued to be not applicable by qualitative researchers, although some kind of qualifying check or measure for their research is stated. If we consider adopting the terms of rigor, and consistency, to determine validity then we mention that the basis of coding and categorization was validated by comparing the kinds of linking words and the attributes with those hosted by a community of experts developing open biomedical ontologies (Ong et al., 2017). Each and every linking word and attribute was compared for its property, and explanation with the ontologies and then only used in the re-representation of the text. The validity of this study was enhanced by utilizing a specific coding scheme and characterizing (Ong et al., 2017). This also led to the development of a Reference Set (see Appendix) of all the relations and attributes that were represented in the content analysis. This way, it not only suggests the rigor in applying the categories, but also facilitates achieving consistency throughout the content analysis process.

Further, it is mentioned that a community of peers can also validate the proposed methodology. Additionally, we also mention that the procedure of content analysis was suggested in preliminary studies that were validated by the international community of peers (Kharatmal and Nagarjuna, 2010, 2016).

7.7 Limitations of the Study

The limitations that apply to qualitative research are mentioned such as time-consuming, labor-intensive work and are mentioned in our study as well. The specific limitations of this study in content analysis have been that it limits to analysis of the cell biology content by choice. However, it is possible that the method can be applied to other topics as well due to the reproducibility of the method.

The study is limited to a specific topic of cell structure and function in biology. The study offers patterns and insights about cell biology. The study also limits in terms of syntax of the English language – subject-predicate-object. We have not worked out the statements with quantifiers and even modalities. This was a choice for selection of form for statements. On a second thought, it is possible to apply the quantifiers – all, some, few, many, etc. as part of the subject and/or object. For example "([all] living beings are composed of cells)"; "([some] cells are non-nucleated)". This can be graphically depicted as ([quantifier]C-LW-[quantifier]C) for propositions shown in Figure 3.3. For applying modalities in the statements, we can use this format as: "([It is possible that] [some] cells are non-nucleated)".

The biological knowledge is usually described in propositional statements for explanation. In addition to this, symbolic representations (equations, graphs), diagrams, pathways, cycles, photographs, maps are often used in textbooks for expression and justification of biological knowledge. Often at school or college level textbooks, use of historical accounts, metaphors and analogies are also part of expression of a phenomena. Use of models, and model-based reasoning as representational tools has a crucial role in representation of biological knowledge. From these diverse aspects of biological knowledge, we have limited to only propositional statements that describe structure and processes.

In order to work out Phase 4, only the process of cell division was considered due to the topic's limitations. Further, any other physiological processes can be mapped using the method. Due to the nature of the research questions, the study is limited to textbook analysis.

The content analysis of biochemical pathways, and texts describing experiments is not yet worked out. Similarly, even mathematical equations are not part of the content analysis, as these are already unambiguous. As our emphasis is on the verbs, we chose to borrow the Set of predicates only.

Biological knowledge shows variations and therefore there exists numerous exceptions. For example, "all living cells have nucleus except mature red blood cells"; " Mammals are usually viviparous except Duck-Billed Platypus", etc. Such statements would require first to represent the fact and then to apply quantifiers, modalities and logical connectives to ascertain the exceptions in the statement. As we think that the initial and preliminary step is to focus on the simple form of of subject-predicate-object statement about descriptions, we selected description based statements, at least at this stage without any rules of exception (to avoid for complicating in initial phases of re-representation). Our study has this aspect as an *exception* while selection of propositional statements for content analysis.

It is possible to extend the method to apply the universal, existential quantifiers, logical connectives, rules of exceptions, modalities to further represent such knowledge. This can be considered to extend our method for future work.

7.8 Implications of the Thesis Work

We now attend to discuss the implications of our work in science education.

Language of Science

While teaching science, the concepts i.e. the class terms are always in focus. We seldom introduce the use of predicates, as these are taken for granted, even though they form meaningful links between concepts. It is known that science tends to be the least ambiguous and the most explicit body of knowledge. Therefore, it is important that students are provided with opportunities to help in the cognitive transition from commonsense language to scientific language, supported by the process of re-representing knowledge structures. In the science education milieu, Halliday (2006) emphasizes teaching scientific language in the early years and not postponing it due to fear of technical words. Along the same thought, Lemke (1990) also suggests to not to be disinclined in scientific language.

Language of Biology

As mentioned the study is limited to cell biology topics. The research work can bear significant implications to study and analyze other relevant topics in biology, for example immunology, microbiology, biochemistry, genetics, etc. Just as the current study has resulted into categories that can identify the nature of cell biology knowledge, it can also provide insights into the nature of other relevant areas of biology knowledge.

Most importantly, we earnestly suggest introducing the students as well as teachers to think about variable properties. We think this would encourage parameterization and quantitative thinking at higher levels. Another strong recommendation from the method is to introduce the learning or representing processes in terms of change, and not just a structure-function relation. The process modeling would not only utilize the variable properties, but also bring about the process-centric nature of biology into focus.

Textbook Analysis

The method of relational content analysis can be applied to the comparison of textbooks as well. Comparative analyses of textbooks have been widely conducted for exploring the domain analysis, knowledge representation, and reasoning requirements for mapping textbooks across the lower level to higher level biology textbooks (Chaudhri et al., 2014d; Koć-Januchta et al., 2020). As a cross-age study, the same domain at the same level can be analyzed for different authors of the text. As a longitudinal study, the same domain can be analyzed with different levels of text. A preliminary comparison of textbooks of the same domain but from different levels indicated that Class 11 text had not only an increasing number of re-represented statements as compared to Class 9 text, but also a larger number of statements depicting processes, and attributes after re-representation. Applying the proximity rating criteria, if the Class 11 text had greater re-represented statements, it implies that it is more proximate to an expert text.

Further, in addition to comparing text, any topic can also be compared across all levels of text. As an example, we compared the topics of nucleus and plasma membrane. For nucleus, the text showed a decreasing number of predicates, while an increasing number of attributes after re-representation, indicating a higher focus on attribution/process based on higher secondary school text. The nominalization of predicates appeared to be more in college level text. For plasma membrane the text showed increasing attributes at all levels of text after re-representation. We found that in two levels of school text, there was a lack (or implicit usage) of attribution before re-representation thereby losing focus on processes. Re-categorization of predicates occurred in two levels of text thereby representing attributes explicitly for increasing the scope of processes. Overall almost all the levels of text have an equal distribution of process-based relations, indicating this is a process-centric topic (unlike structure-based topic). The topic of the nucleus seemed more process-centric when compared with the topic of the plasma membrane which seemed structure-based mostly. The transition of the nucleus topic from structure-based to process-based occurs with increasing levels of text.

Textbook Re-Writing

We suggest the implication of the study for re-representing the textbooks. Biology is considerd to be mostly a descriptive subject. We provide the Reference Set that can be utilized during textbook writing by the experts or committees. Wherever possible, the re-represented concept maps can be complemented with the passages or can be considered for assignments.

Meaningful Learning

The rewritten science text can be represented in a graphical format as well, such as concept mapping. These have been widely used as an effective tool for organizing knowledge, students' understanding, knowledge restructuring, assessment, etc. (Novak and Gowin, 1984). In concept mapping, concepts are connected to other concepts using linking words for meaningful learning. It is mentioned that the linking words when are semantically well-defined can create unambiguous maps for meaningful learning (Kharatmal and Nagarjuna, 2006b).

Assessment Tool

The concept mapping has been widely used as a pedagogical and assessment tool applying the scoring method (Novak and Gowin, 1984; Mintzes et al., 2000). According to it, each proposition is scored if two concepts are linked with a linking word. This seems graphical, as it does not mention to the use of well-defined linking words for disambiguation or consistency. As the re-representation method focuses on the 'semantics' of the linking words, we suggest adding this criterion with higher points. Considering the focus on linking words, concept maps with fill-in-concepts, fill-in-linking-words are suggested, so that it can facilitate thinking about meaningful relations. These maps with a focus on well-defined linking words can help to investigate misconceptions, to develop science vocabulary, etc.

Science in English as a Second Language

In the language and educational contexts, it is known that the number of people who learn science through English as a Second Language (ESL) is larger than the number of people who learn science in English as a First Language (EFL). When English is not the native language, the learning can happen explicitly. This could be because the ambiguity of "has',"is-a', etc. may not occur. We suggest that when text is written as nuclear predicates, then science text can become accessible to ESL learners.

Another important application of the nuclear predicate would be in the area of Second Language Acquisition (SLA). The First Language Acquisition (FLA) i.e. the mother tongue, occurs naturally. However, during the SLA, learners face difficulties that depend on various factors. SLA is a conscious effort of learning its lexicon, grammar, etc. (Ellis, 1997). While it is considered that a Universal Grammar (UG) is associated with FLA, a unique abstract linguistic system of grammatical rules called the interlanguage (IL) is associated with SLA (Ellis, 1997). To develop this IL, the SLA learners make use of learning strategies, with the appropriate exposure to a second language (L2). We suggest that the use of nuclear predicates can facilitate one such learning strategy. Due to their domain-general ability, these can be used in SLA for learners to learn the domain-specific class terms that are available in the L2. This would help in schools where English is a second language for learning science. This can benefit pedagogy in science classrooms and language classrooms.

Controlled Vocabulary of Predicates

We finally discuss the application of our method for creating a controlled vocabulary of nuclear predicates. The understanding of the meaning of terms is associated with developing a deep vocabulary knowledge (Schmitt, 2007). Employing metacognitive strategies by forming connections (meaningful links) between terms can help students in learning biology vocabulary (Gu and Johnson, 1996). Creating a controlled vocabulary of biology can reduce the burden of learning the heavy jargon for biology students (Bennett et al., 2013). Towards this goal, Chaudhri et al. (2019, p. 131) are creating intelligent, interactive textbooks by creating a glossary page of explicit statements of universal truths. The glossary is arranged using relation terms as metadata for explicit statements applying the method of Aristotelian definition. In their illustration of a glossary of one such term about eukaryotic cells, the predicates being used are: "has part" as relations and "size", "diameter' as variable properties. Our work of semantic content analysis along with the compiled list of relations and attributes can facilitate as a precursor to creating such a knowledge base.

After re-representation, the passages express the same content knowledge. All the predicates and the class terms used in the rewritten passages are available in the published ontologies of biology. The re-represented passage are very close to, if not identical to, a controlled natural language (Fuchs et al., 2008; Kuhn, 2010), controlled vocabulary of biology (Bennett et al., 2013). The process preserves both meaning and truth value, as well as the grammar, but makes it more accessible for non-native speakers and machines. Moreover, by making the implicit meaning explicit, we can achieve rigor (Kharatmal and Nagarjuna, 2010; Smith et al., 2005). It may be noted that the re-re-represented passages can also be graphically represented in the form of concept graphs, concept maps, knowledge maps, semantic networks, and ontology graphs. (Novak, 1990; Fisher et al., 2000; Chaudhri et al., 2021).

Automated Method

For the current research study, the work of all the data representation, and content analysis was carried out manually. It is possible that the work can be automated. Applying the concept mapping tool, semantic network technologies, and graphics-based software packages, the entire work of representation of text and re-representation can be automated. This exercise even can help in developing teaching, learning, and assessment tools in education. We envisage, that teachers, students, and researchers will be benefited from the method developed. Our method can be also used to represent textbooks in teaching learning exercises and assignments.

The concept map propositions in the form of triples can be extended to create concept lattices (Ganter, 1997; Priss, 2010). By focusing on relations and attributes, concept neighborhood lattices of dependencies, associations and of dynamic propositions can be generated (Kharatmal, 2014).

Researchers in KR, have conducted comparative analysis studies using life science textbooks (Chaudhri et al., 2014b). We too represent the structure based relations as "has part", and spatial entity relations as "has region". Their study applies reasoning methods and uses this to create template for creating questions and answers. While our method as of now is just as a representation stage, the future work can involve about introducing reasoning, question-answering methods.

7.9 Recommendations

Pedagogical Tool

The re-representation method can be used as a pedagogical tool. An empirical research study can be conducted to find its effectiveness in teaching and learning biology, where it can have control groups with standard teaching, standard concept mapping, and an experimental group of re-represented concept mapping in the classroom. Through collaborative concept mapping teachers can create and use for lesson development, lesson delivery, assessment, pre-post evaluation, etc.

Model based reasoning

The state change model for process representation can be used in parameterization in models for scientific experiments. Science employs conceptual structures (Gardenfors, 2000) for a given dimension, and the data properties that are applied in each conceptual structure are generic, i.e. the same data properties in the same dimension are being applied for modeling. It is a widely accepted view of science, both among the practitioners and also by the philosophers of science that scientific knowledge uses models for describing and explaining phenomena (Nersessian, 1998; Magnani et al., 1999). Modeling a phenomenon involves describing the phenomena using variables, which in ontological language are attributes (data properties). A modeling language along with its semantics and constraints reflects the structure of biological phenomena it represents (Hoehndorf et al., 2011). We specifically emphasize the role of attributes i.e. data properties for model-based reasoning and quantitative reasoning in biology. It is necessary to represent knowledge of structure and function explicitly in a model which can further be used for model-based reasoning using logical principles.

Knowledge Representation in Education

The relations and attributes extracted from the content analysis can be used for creating ontology graphs, intelligent textbooks, querying, designing question-answering programs, etc. (Barker et al., 2004; Chaudhri et al., 2013). In recent research, an Automated Relation Extractor (ARE) method was used to create ontology graphs for KB Bio 101, using 5 categories of relations – *taxonomic, structural, function, event, causal* (Chaudhri et al., 2021). Our work involved a fully manual method for relation extraction from biology text arriving at 6 categories of relations and 3 categories of attributes, some of which are common with the ARE work.

We see parallels in SRI's (Chaudhri, 2014) library of resources on concept graphs. The researchers uses concepts linked with labeled linkages and depict distinction between structure terms, process terms and attributes. We also follow these distinctions in our re-representation method. The common part of this concept graph and our re-representation method is: structure is linked to another structure using the "has-part" relation; structure is linked with process term using the "has-function" relation. The process terms in our method are nominalized and linked with "has role/has function" relation. In their concept graph, each process term is depicted as a term. In some places, the attribute term is also used as relation term. The resource is useful for future work.

LLMs and chatGPT

The recent boom of using language learning models (LLMs) for example the chatGPT, has created concerns in various domains. The LLMs are machine learning models providing with human-like responses to a prompt. The LLMs are trained on the massive corpus of online text, learning about the patterns and links between words. Although their knowledge base is vast, it is also sometimes incorrect and biased, and moreover, it does not explicitly understand the semantics or relationships in that content. (A test prompt in chatGPT indicates this confusion: `https://chat.openai.com/chat/486ca0a3-963e-4be3-8a6f-559c3b8ff5eb`).

On the contrary, ontologies are representations of the domain in terms of concepts, relationships, properties, axioms and rules providing a framework for a deeper understanding of that domain. Ontologies are used to enable machine reasoning and semantic understanding, allowing a system to draw inferences and to derive new information and relationships between entities (Ontology-Summit, 2023a,b). Our work follows the published and validated ontologies, generic library of resources to create the Reference Set.

7.10 Future Scope

The future work can be in the direction of using the method for: creating intelligent textbooks, teaching-learning focusing on linking words, comparing novice's and expert's language, creating question-answering programs, disambiguation of text, modeling of experiments, creating interactive textbooks, annotating the science text focusing on the nuclear predicates, adding criteria in the intelligent grammar correcting applications for academic text, designing games that facilitate the use of writing text with 'nuclear predicates', etc. It may even become a criterion for the assessment of the written text.

There is scope for using the model to represent the dynamics of biology knowledge and extend it beyond processes, in terms of metabolic pathways, various biogeochemical cycles. The method can also be used for comparing the nature of knowledge within biological sciences and other physical sciences.

As future scope of extending the research work, the learning of the Re-representation method as technique can be imparted to teachers by conducting workshops. Pedagogical interventions and strategies can also be developed for introducing the method in collaboration with teachers to engage with students. The effectiveness of the method in teaching-learning can be studied as an intervention with pre-post studies by conducting empirical studies. These ideas can culminate into developing a research program for continuous engagements and research areas that are yet to explored as an offshoot from this thesis work.

Another interesting area as a future scope of the study could be for venturing to apply the Re-representation method in a bi-lingual medium of instruction for science, or semi-English medium of instruction. In most educational settings, the language of science is dominated by English language. However considering the diverse and multilingual settings in our country, it is possible to explore the method for representing science in Marathi and Hindi languages, in which the syntax is totally different from English language. This work certainly would be a challenge and may bring in additional novelty to the research work.

7.11 Reflections

Tracing the journey of this exploratory study, we would like to reflect on the process and the findings of the research work—a model for representing cell biology knowledge. The Four ontological categories – that are sufficient to represent all of cell biology knowledge in terms of structure and dynamics – have been a significant outcome. The scope for quantification of biology by making the distinction between relations and attributes and further for representing processes as change has been a unique feature. In terms of the language of science within the school of SFL, extending the framework for experimentation is also an added achievement of the study.

We agree that it was time-consuming as this is a developmental process and initially takes time. The entire research of developing and using the method for content analysis has been a manual task with iterative refinements. It is possible that eventually, with practice and continuous focus on clarity in expression, the method can be learned. One of the points about the nature of linking words or relations is that these belong to the natural language vocabulary and are available in any everyday experience. The method is about the explicit usage of linking words. At the same time, we agree that with implicit linking words, the meaning still gets conveyed; however, we propose that with repeated training and its usage (or habit), this procedure may also eventually become implicit and can be used seamlessly. It may be the initial inertia at first, but later on, it can be ingrained in thinking.

We agree that the context can provide meaning. But it may not necessarily compromise the clarity of expression. On the other hand, the context helps to think about the dimensions of the statements. Therefore, in order to make this into unambiguous statements, well-defined relations can become helpful. We are suggesting this method for science communication and expression of statements. On the contrary, the ambiguities and contexts are part of poetry and have scope for various interpretations, unlike scientific knowledge.

Some lessons learned during the process as a researcher include perseverance in designing and developing a model. Engaging in an interdisciplinary aptitude and having a bird's eye view while creating a model was advantageous. Even though this research work is in the area of science education, we were able to get inspired by the insights from philosophy, cognitive science, linguistics, knowledge representation, semantic web, computer science, logic, etc.

We hope the Re-representation method focusing on semantics may be explored in practicum for the language of science and science communication. In this exploratory study, we think we have succeeded in convincing the reader about the role of predicates in explicating meaning and striving toward rigor. We hope the study helped in bringing the significance of predicates to the foreground for a study of semantics in the language of science.

"I link, therefore I am".

Bibliography

OWL - Semantic Web Standards, 2012. URL https://www.w3.org/OWL/.

Agustín Adúriz-Bravo, Andrea Revel Chion, and Alejandro P. Pujalte. Scientific language. In Richard Gunstone, editor, *Encyclopedia of Science Education*, pages 1–4. Springer Netherlands, Dordrecht, 2021. ISBN 978-94-007-6165-0. doi: 10.1007/978-94-007-6165-0_262-1. URL https://doi.org/10.1007/978-94-007-6165-0_262-1.

Terry Anderson and Julie Shattuck. Design-Based Research: A Decade of Progress in Education Research? *Educational Researcher*, 41(1):16–25, jan 2012. ISSN 0013-189X. doi: 10.3102/0013189X11428813. URL https://journals.sagepub.com/doi/abs/10.3102/0013189X11428813. Publisher: American Educational Research Association.

R. Arp and B. Smith. Function, role, and disposition in basic formal ontology. *Nature Preceedings*, pages 1–4, 2008b. URL https://www.nature.com/articles/npre.2008.1941.1.

David Ausubel, Joeseph Novak, and Holt Hanesian. *Educational psychology: a cognitive view*. Holt, Rinehart and Winston, 1978. ISBN 9780030899515.

Riccardo Baratella. What are we debating about when we debate about processes and events? JOWO volume on CEUR-WS.org, 2020. URL https://foust2020.inf.unibz.it/wp-content/uploads/2020/09/foust1.pdf.

Ken Barker, Bruce Porter, and Peter Clark. A library of generic concepts for composing knowledge bases. In *Proceedings of the 1st International conference on Knowledge capture*, pages 14–21. Association for Computing Machinery, New York, United States, 2001. URL https://www.ai.sri.com/~halo/public/clib/20130328/components/specs/slotdictionary.html.

Ken Barker, Vinay K. Chaudhri, Shaw Yi Chaw, Peter Clark, James Fan, David J. Israel, Sunil Mishra, Bruce W. Porter, Pedro Romero, Dan Tecuci, and Peter Z. Yeh. A Question-Answering System for AP Chemistry: Assessing KR&R Technologies. In Didier Dubois, Christopher A. Welty, and Mary-Anne Williams, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference (KR2004), Whistler, Canada, June 2-5, 2004*, pages 488–497. AAAI Press, 2004. URL http://www.cs.utexas.edu/users/ai-lab?kr04-shaken.

Lucas Bechberger and Kai-Uwe Kühnberger. A thorough formalization of conceptual spaces. In *KI 2017: Advances in Artificial Intelligence*, pages 58–71. Springer International Publishing, 2017. doi: 10.1007/978-3-319-67190-1_5. URL https://doi.org/10.1007/978-3-319-67190-1_5.

Sean Bechhofer. *OWL: Web Ontology Language*, pages 2008–2009. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_1073. URL https://doi.org/10.1007/978-0-387-39940-9_1073.

Mariette Bengtsson. How to plan and perform a qualitative study using content analysis. *NursingPlus Open*, 2:8–14, January 2016. ISSN 2352-9008. doi: 10.1016/j.npls.2016.01.001. URL https://www.sciencedirect.com/science/article/pii/S2352900816000029.

Brandon Bennett, Vinay K. Chaudhri, and Nikhil Dinesh. A Vocabulary of Topological and Containment Relations for a Practical Biological Ontology. In Thora Tenbrink, John G. Stell, Antony Galton, and Zena Wood, editors, *Spatial Information Theory - 11th International Conference, COSIT 2013, Scarborough, UK, September 2-6, 2013. Proceedings*, volume 8116 of *Lecture Notes in Computer Science*, pages 418–437. Springer, 2013. URL https://doi.org/10.1007/978-3-319-01790-7_23.

Olivier Bodenreider, Joyce A. Mitchell, and Alexa T. McCray. Biomedical Ontologies. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 76–78, 2005. ISSN 2335-6936. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4300097/.

Patricia Boechler, Mary Ingraham, Luis Marin, Brenda Dalen, and Erik Dejong. Making the Implicit Explicit: Music Listening, Blended Delivery and the Representational Redescription Model. *International Journal for Cross-Disciplinary Subjects in Education*, 6:2095–2105, March 2015. doi: 10.20533/ijcdse.2042.6364.2015.0291.

Edward De Bono. *Lateral Thinking: A Textbook of Creativity*. Penguin Books, 1977. ISBN 978-0-14-021978-4.

Ronald J. Brachman. What is-a is and isn't: an analysis of taxonomic links in semantic networks. *IEEE Computer*, 16(10):30–36, 1983. URL https://ieeexplore.ieee.org/document/1654194.

Amy G. Briggs, Stephanie K. Morgan, Seth K. Sanderson, Molly C. Schulting, and Laramie J. Wieseman. Tracking the resolution of student misconceptions about the central dogma of molecular biology. 17(3):339–350, 2016. doi: 10.1128/jmbe.v17i3.1165. URL https://journals.asm.org/doi/10.1128/jmbe.v17i3.1165. Publisher: American Society for Microbiology.

Ann L. Brown. Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *Journal of the Learning Sciences*, 2(2):141–178, apr 1992. ISSN 1050-8406. doi: 10.1207/s15327809jls0202_2. Publisher: Routledge.

Patryk Burek, Robert Hoehndorf, Frank Loebe, Johann Visagie, Heinrich Herre, and Janet Kelso. A top-level ontology of functions and its application in the open biomedical ontologies. *Bioinformatics (Oxford, England)*, 22(14):e66–e73, jul 2006. ISSN 1367-4811. doi: 10.1093/bioinformatics/btl266. URL https://doi.org/10.1093/bioinformatics/btl266.

Cathal Butler. Evaluating the Utility and Validity of the Representational Redescription Model as a General Model for Cognitive Development. October 2008.

Neil A. Campbell and Jane B. Reece. *Biology*. Pearson, Benjamin Cummings, San Fancisco, USA, 8th edition, 2008. ISBN 9780321536167.

Alberto J. Canas and Marco Carvalho. Concept maps and AI: an unlikely marriage? In *Revista Brasileira de Informáticana Educação*, 2004. URL Canas-CmapsAI-Sbie2004.pdf.

Alberto J Canas and Joeseph D Novak. Re-examining the foundations for effective use of concept maps. In *Concept maps: theory, methodology, technology. Proceedings of the second international conference on concept mapping*, volume 1, pages 494–502, 2006.

Alberto J. Canas, Joseph D. Novak, and Priit Reiska. How good is my concept map? am i a good cmapper? 7(1):6, 2015. Publisher: The University of Hong Kong-Faculty of Education.

Alberto J. Canas, Priit Reiska, and Aet Möllits. Developing higher-order thinking skills with concept mapping: A case of pedagogic frailty. 9(3):348, 2017. URL https://doi.org/10.34105/j.kmel.2017.09.021. Publisher: The University of Hong Kong-Faculty of Education.

Antonella Carassa and Maurizio Tirassa. Representational redescription and cognitive architectures. 17, 1994. doi: 10.1017/S0140525X00036670.

Kathleen Carley. Content Analysis. In R. E. Asher and et al., editors, *The Encyclopedia of Language and Linguistics*, volume 2, pages 725–730. Pergamon Vpress, Endinburgh, UK, 1990. URL http://www.casos.cs.cmu.edu/publications/papers/carley_1990_contentanalysis.PDF.

Rudolf Carnap. Empiricism and the language of science. *Synthese*, pages 33–35, 1938.

Marci Catanzaro. Using Qualitative Analytical Techniques. In Nancy Woods and Marci Catanzaro, editors, *Nursing Research: Theory and Practice*. Mosby Incorporated, St Louis, 1988.

Alberto J. Cañas and Joseph D. Novak. Concept mapping using Cmap tools to enhance meaningful learning. In *Knowledge Cartography*, pages 25–46. Springer, 2008. URL https://link.springer.com/chapter/10.1007/978-1-84800-149-7_2.

CBSE. CBSE, 1962. URL https://www.cbse.gov.in/cbsenew/cbse.html.

Vinay Chaudhri. SRI library of concept graphs, 2014. URL https://www.ai.sri.com/~halo/public/exported-kb/graphs/page.html.

Vinay Chaudhri, Nikhil Dinesh, and Daniela Inclezan. Creating a knowledge base to enable explanation, reasoning, and dialog: Three lessons. 3:183–200, 2014a. URL http://www.cogsys.org/journal/volume3/article-3-12.pdf.

Vinay Chaudhri, Daniel Elenius, Andrew Goldenkranz, Allison Gong, Maryann Martone, William Webb, and Neil Yorke-Smith. Comparative analysis of knowledge representation and reasoning requirements across a range of life sciences textbooks. 5(1): 1–19, 2014b. URL https://jbiomedsem.biomedcentral.com/articles/10.1186/2041-1480-5-51.

Vinay K. Chaudhri and Daniela Inclezan. Representing States in a Biology Textbook. In *2015 AAAI Spring Symposia, Stanford University, Palo Alto, California, USA,* *March 22-25, 2015*. AAAI Press, 2015. URL https://cdn.aaai.org/ocs/10328/10328-45257-1-PB.pdf.

Vinay K. Chaudhri, Andrew Goldenkranz, Richard Fikes, and Patrice Seyed. What is hard about representing biology textbook knowledge. In Mark A. Musen and Óscar Corcho, editors, *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP 2011), June 26-29, 2011, Banff, Alberta, Canada*, pages 185–186. ACM, 2011. doi: 10.1145/1999676.1999717. URL https://dl.acm.org/doi/abs/10.1145/1999676.1999717.

Vinay K. Chaudhri, Britte Cheng, Adam Overtholtzer, Jeremy Roschelle, Aaron Spaulding, Peter Clark, Mark Greaves, and Dave Gunning. Inquire biology: A textbook that answers questions. 34(3):55–72, 2013. URL https://onlinelibrary.wiley.com/doi/10.1609/aimag.v34i3.2486.

Vinay K. Chaudhri, Nikhil Dinesh, and Stijn Heymans. Conceptual Models of Energy Transfer and Regulation. In Pawel Garbacz and Oliver Kutz, editors, *Formal Ontology in Information Systems - Proceedings of the Eighth International Conference, FOIS 2014, September, 22-25, 2014, Rio de Janeiro, Brazil*, volume 267 of *Frontiers in Artificial Intelligence and Applications*, pages 263–276. IOS Press, 2014c. doi: 10.3233/978-1-61499-438-1-263. URL https://ebooks.iospress.nl/publication/37975.

Vinay K. Chaudhri, Daniel Elenius, Sue Hinojoza, and Michael A. Wessel. KB_Bio_101: Content and Challenges. In Pawel Garbacz and Oliver Kutz, editors, *Formal Ontology in Information Systems - Proceedings of the Eighth International Conference, FOIS 2014, September, 22-25, 2014, Rio de Janeiro, Brazil*, volume 267 of *Frontiers in Artificial Intelligence and Applications*, pages 415–420. IOS Press, 2014d. doi: 10.3233/978-1-61499-438-1-415. URL https://www.sri.com/wp-content/uploads/2021/12/1981.pdf.

Vinay K. Chaudhri, Yan Gong, Craig Heller, and Shizuka Yamada. *Explicating the Logic of Biology to Support Critical Thinking*, pages 127–134. Springer International Publishing, Cham, 2019. ISBN 978-3-030-26578-6. doi: 10.1007/978-3-030-26578-6_10. URL https://link.springer.com/chapter/10.1007/978-3-030-26578-6_10.

Vinay K. Chaudhri, Matthew Boggess, Han Lin Aung, Debshila Basu Mallick, Andrew C. Waters, and Richard G. Baraniuk. A Case Study in Bootstrapping Ontology Graphs from Textbooks. In Danqi Chen, Jonathan Berant, Andrew McCallum, and Sameer Singh, editors, *3rd Conference on Automated Knowledge Base Construction, AKBC 2021, Virtual, October 4-8, 2021*, 2021. doi: 10.24432/C58C7T. URL https://openreview.net/pdf?id=nDe2D8DDXKR.

Michelene T. H. Chi and Rod D. Roscoe. *The Processes and Challenges of Conceptual Change*, pages 3–27. Springer Netherlands, Dordrecht, 2002. ISBN 978-0-306-47637-2. doi: 10.1007/0-306-47637-1_1. URL https://doi.org/10.1007/0-306-47637-1_1.

Eugene L. Chiappetta, David A. Fillman, and Godrej H. Sethna. Procedures for conducting content analysis of science textbooks. Technical report, Houston, TX: University of Houston, Department of Curriculum and Instruction., 2004.

Paul M. Churchland. *Matter and Consciousness: A Contemporary Introduction to the*

Philosophy of Mind. MIT Press, 1988. ISBN 9780262530743. URL https://mitpress.mit.edu/9780262530743/matter-and-consciousness/.

Andy Clark and Annette Karmiloff-Smith. The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind & Language*, 8 (4):487–519, 1993. ISSN 1468-0017. doi: 10.1111/j.1468-0017.1993.tb00299.x. Place: United Kingdom Publisher: Blackwell Publishing.

Peter Clark, William R. Murray, Phil Harrison, and John Thompson. Naturalness vs. predictability: A key debate in controlled languages. In Norbert E. Fuchs, editor, *Controlled Natural Language*, pages 65–81, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-14418-9. URL https://link.springer.com/chapter/10.1007/978-3-642-14418-9_5.

Louis Cohen, Lawrence Manion, and Keith Morrison. *Research Methods in Education.* Routledge, New York, seventh edition. edition, 2015.

Louis Cohen, Lawrence Manion, and Keith Morrison. *Research Methods in Education.* Routledge, London, eight edition, October 2017. ISBN 978-1-315-45653-9. doi: 10.4324/9781315456539.

Allan Collins, Diana Joseph, and Katerine Bielaczyc. Design Research: Theoretical and Methodological Issues. *Journal of the Learning Sciences*, 13(1):15–42, jan 2004. ISSN 1050-8406. doi: 10.1207/s15327809jls1301_2.

Irving M. Copi, Carl Cohen, and Kenneth D. McMahon. *Introduction to logic.* Macmillan New York, 1972. URL http://www.stoqatpul.org/lat/materials/copi_cohen_8ed.pdf.

Julio V. Costa, Francisco E. Rocha, and Eloi L. Favero. Linking phrases in concept maps: A study on the nature of inclusivity. In A. J. Cañas, J. D. Novak, and F. M. González, editors, *Concept Maps: Theory, Methodology, Technology.*, 2004.

John W. Creswell. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches: 0.* SAGE Publications, Inc, Thousand Oaks, Calif, third edition edition, July 2008. ISBN 978-1-4129-6557-6.

John W. Creswell. *Qualitative inquiry & research design: Choosing among five approaches.* Thousand Oaks, CA, 4th edition, 2012.

Sarah Critten, Karen Pine, and Dorothy Steffler. Spelling development in young children: A case of representational redescription? *Journal of Educational Psychology*, 99:207–220, 2007. ISSN 1939-2176. doi: 10.1037/0022-0663.99.1.207. Place: US Publisher: American Psychological Association.

Deirdre Davies and Jenny Dodd. Qualitative Research and the Question of Rigor. *Qualitative Health Research*, 12(2), 2002. doi: https://doi.org/10.1177/104973230201200211. URL https://journals.sagepub.com/doi/10.1177/104973230201200211.

Kirsten E. de Ries, Harmen Schaap, Anne-Marieke M. J. A. P. van Loon, Marijke M. H. Kral, and Paulien C. Meijer. A literature review of open-ended concept maps as a research instrument to study knowledge and learning. *Quality & Quantity*, 56(1): 73–107, February 2022. ISSN 1573-7845. doi: 10.1007/s11135-021-01113-x. URL https://doi.org/10.1007/s11135-021-01113-x.

Eduardo D. P. De Robertis and Edward M. F. De Robertis. *Cell and Molecular Biology*. Lea & Febiger, New Delhi, India, 1985. URL https://archive.org/details/dli.scoerat.2286cellandmlecularbiology.

Norman K. Denzin and Yvonna S. Lincoln, editors. *The landscape of qualitative research: Theories and issues.* Thousand Oaks: Sage Publications., CA, 1998.

Natalia Derbentseva, Frank Safayeni, and Alberto J. Cañas. Concept maps: Experiments on dynamic thinking. *Journal of Research in Science Teaching*, 44(3):448–465, 2007. ISSN 1098-2736. doi: 10.1002/tea.20153. URL http://onlinelibrary.wiley.com/doi/10.1002/tea.20153/abstract.

Yaegan Doran and James Martin. Field relations: Understanding scientific explanations. In Karl Maton, James R. Martin, and Yaegan J. Doran, editors, *Teaching Science: Language, Knowledge, Pedagogy*. Routledge, 2021.

Dov Dori. *Object-Process Methodology*. Springer-Verlag, New York, 1 edition, 2002. ISBN 9783540654711. URL https://link.springer.com/book/10.1007/978-3-642-56209-9.

Barbara Downe-Wamboldt. Content analysis: Method, applications, and issues. *Health Care for Women International*, 13(3):313–321, January 1992. ISSN 0739-9332. doi: 10.1080/07399339209516006. URL https://doi.org/10.1080/07399339209516006. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07399339209516006.

Rod Ellis. *Second Language Acquisition*. Oxford University Press, Oxford, 1997. URL https://www.jstor.org/stable/44487448.

Esra Erdem and Reyyan Yeniterzi. Transforming Controlled Natural Language Biomedical Queries into Answer Set Programs. In *BioNLP@HLT-NAACL*, 2009. doi: 10.3115/1572364.1572381. URL https://aclanthology.org/W09-1315.pdf.

Maria Evagorou and Jonathan Osborne. The role of language in the learning and teaching of science. In *Good Practice in Science Teaching: What Research has to Say*. Open University Press, 2010.

Christian Fellbaum. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998. URL https://direct.mit.edu/books/book/1928/WordNetAn-Electronic-Lexical-Database.

Kathleen M. Fisher. Semantic networking: The new kid on the block. *Journal of Research in Science Teaching*, 27(10):1001–1018, 1990. ISSN 1098-2736. URL https://doi.org/10.1002/tea.3660271008.

Kathleen M. Fisher. Overview of Knowledge Mapping. In *Mapping biology knowledge*, pages 5–23. Springer, 2002. URL https://link.springer.com/book/10.1007/0-306-47225-2.

Kathleen M. Fisher and Michael R. Kibby, editors. *Knowledge Acquisition, Organization, and Use in Biology*. Springer, 1 edition, February 1996. ISBN 3540604154. URL https://link.springer.com/book/10.1007/978-3-642-61047-9.

Kathleen M Fisher, James H Wandersee, and David E Moody. *Mapping biology knowl-*

edge, volume 11. Springer, 2000. URL https://link.springer.com/book/10.1007/0-306-47225-2.

Jerry Fodor. *Modularity of mind: An essay on faculty psychology.* MIT Press, Cambridge, 1983.

Marta Fonseca, Beatriz Oliveira, Inês Canha, Hélder Dores, Madalena Pinheiro Santos, Vasco Cremon Lemos, Alexandra Verdasca, Mariana Branco, Ana Rita Póvoa, Pedro Carreiro-Martins, António Rendas, and Nuno Neuparth. Analyzing the use of linking words in concept maps designed for pathophysiology learning in medicine. In *7th International Conference on Higher Education Advances (HEAd'21)*, pages 1–8. Universitat Politècnica de València, 2021. doi: 10.4995/HEAd21.2021.13036. URL http://ocs.editorial.upv.es/index.php/HEAD/HEAd21/paper/view/13036.

Natalya Fridman Noy and Carole D. Hafner. Ontological foundations for experimental science knowledge bases. *Applied Artificial Intelligence*, 14(6):565–618, July 2000. ISSN 0883-9514. doi: 10.1080/08839510050076972.

Norbert E. Fuchs. Understanding Texts in Attempto Controlled English. *Controlled Natural Language*, pages 75–84, 2018. doi: 10.3233/978-1-61499-904-1-75. URL http://attempto.ifi.uzh.ch/site/pubs/papers/cnl2018main_fuchs.pdf.

Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. Attempto controlled english for knowledge representation. In *Reasoning web*, pages 104–124. Springer, 2008. URL https://link.springer.com/chapter/10.1007/978-3-540-85658-0_3.

Rudolf Ganter, Bernhard Wille. Applied lattice theory: Formal concept analysis. In G. Gratzer and Birkhauser, editors, *General Lattice Theory.*, Lecture Notes in Computer Science. Preprints, 1997. URL http://www.math.tu-dresden.de/$\sim$ganter/psfiles/concept.ps.

Peter Gardenfors. *Conceptual Spaces: The Geometry of Thought.* MIT Press, USA, 2000. ISBN 978-0-262-57219-4. URL https://direct.mit.edu/books/book/2532/Conceptual-SpacesThe-Geometry-of-Thought.

Peter Gardenfors and Massimo Warglien. Using Conceptual Spaces to Model Actions and Events. *Journal of Semantics*, 29(4):487–519, 2012. doi: 10.1093/jos/ffs007. URL https://doi.org/10.1093/jos/ffs007.

Paul L Gardner. Language difficulties of science students. *Australian Science Teachers Journal*, 20(1):63–76, 1974.

Ronald N. Giere. *Explaining science: A cognitive approach.* University of Chicago Press, 1990. URL https://press.uchicago.edu/ucp/books/book/chicago/E/bo3622319.html.

Ashok K. Goel, Spencer Rugaber, and Swaroop Vattam. Structure, behavior, and function of complex systems: The structure, behavior, and function modeling language. *AI EDAM*, 23(1):23–35, February 2009. ISSN 1469-1760, 0890-0604. doi: 10.1017/S0890060409000080.

Nahid Golafshani. Understanding Reliability and Validity in Qualitative Research. *The Qualitative Report*, 8(4):597–606, December 2003. ISSN 1052-0147. doi: 10.46743/2160-3715/2003.1870. URL https://nsuworks.nova.edu/tqr/vol8/iss4/6.

Ulla H Graneheim and Berit Lundman. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today*, 24(2):105–112, February 2004. ISSN 0260-6917. doi: 10.1016/j.nedt.2003.10.001. URL `http://www.sciencedirect.com/science/article/pii/S0260691703001515`.

Yongqi Gu and Robert K. Johnson. Vocabulary learning strategies and language learning. In *Language and Learning*. Wiley International., 1996. URL `https://onlinelibrary.wiley.com/doi/10.1111/j.1467-1770.1996.tb01355.x`.

Nicola Guarino, Riccardo Baratella, and Giancarlo Guizzardi. Events, their names, and their synchronic structure. *Applied Ontology*, 17(2):249–283, January 2022. ISSN 1570-5838. doi: 10.3233/AO-220261.

Giancarlo Guizzardi and Terry Halpin. Ontological foundations for conceptual modelling. *Applied Ontology*, 3:1–12, 01 2008. doi: 10.3233/AO-2008-0046.

Peter Gärdenfors. *Conceptual spaces: The geometry of thought*. MIT press, 2004. URL `https://mitpress.mit.edu/9780262572194/conceptual-spaces/`.

Peter Gärdenfors. Semantics based on conceptual spaces. In *Indian Conference on Logic and Its Applications*, pages 1–11. Springer, 2011. URL `https://link.springer.com/chapter/10.1007/978-3-642-18026-2_1`.

Carole D. Hafner and Natalya. Fridman. Ontological foundations for biology knowledge models. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, 4:78–87, 1996. ISSN 1553-0833.

Michael A. K. Halliday. *An Introduction to Functional Grammar*. Edward Arnold., London, 1st edition, 1985.

Michael A. K. Halliday. *The Language of Science*. Continuum International Publishing Group, London, 2006. ISBN 9780826488275.

Michael A. K. Halliday and James R. Martin. *Writing Science: Literacy and Discursive Power*. The Falmer Press, London, 1993. ISBN 9780750700467.

Michael A. K. Halliday and Christian M. I. Matthiessen. *An Introduction to Functional Grammar*. Routledge, London, 2004.

Michael A. K. Halliday and Christian M.I.M. Matthiessen. *Halliday's Introduction to Functional Grammar*. Routledge, UK, 2014.

Rom Harré. *Principles of scientific thinking*. University of Chicago Press, Chicago, 1970.

David Hay, Ian Kinchin, and Simon Lygo-Baker. Making learning visible: the role of concept mapping in higher education. 33(3):295–311, 2008. ISSN 0307-5079. URL `https://doi.org/10.1080/03075070802049251`. Publisher: Routledge _eprint: https://doi.org/10.1080/03075070802049251.

Patrick Hayes. RDF semantics, 2004. URL `https://www.w3.org/TR/2004/REC-rdf-mt-20040210/`.

Robert Hoehndorf, Michel Dumontier, John H Gennari, Sarala Wimalaratne, Bernard de Bono, Daniel L Cook, and Georgios V Gkoutos. Integrating systems biology models and biomedical ontologies. *BMC Systems Biology*, 5:124, August 2011. ISSN 1752-0509. doi: 10.1186/1752-0509-5-124.

Robert Hoehndorf, Paul N. Schofield, and Georgios V. Gkoutos. The role of ontologies in biological and biomedical research: a functional perspective. *Briefings in Bioinformatics*, 16(6):1069–1080, November 2015. ISSN 1467-5463. doi: 10.1093/bib/bbv011.

Marko Hofmann. *Ontologies in Modeling and Simulation: An Epistemological Perspective*, pages 59–87. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-31140-6. doi: 10.1007/978-3-642-31140-6_3. URL https://doi.org/10.1007/978-3-642-31140-6_3.

James Hollan, Edwin Hutchins, and David Kirsh. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 7(2):174–196, 2000. Publisher: ACM New York, NY, USA.

Quan-Thanh Huynh and Yu-Chuan Yang. Impact of fill-in-the-nodes concept maps on low prior-knowledge students learning chemistry: a study on the learning achievements and attitude toward concept maps. 25(1):360–374, 2024. URL https://pubs.rsc.org/en/content/articlelanding/2024/rp/d3rp00238a.

IHMC. CmapTools, 2011. URL http://cmap.ihmc.us.

Robert Javonillo and Catherine Martin-Dunlop. Linking phrases for concept mapping in introductory college biology. 45(3), 2019.

Ingvar Johansson. Four Kinds of Is_a Relation. In *Applied Ontology: An Introduction*, pages 235–254. De Gruyter, May 2013. ISBN 9783110324860.

Tamara M. Juznic. A contrastive study of nominalization in the systemic functional framework. *Languages in Contrast*, 12(2):251–276, 2012. doi: 10.1075/lic.12.2.08mik.

Gary E. Kaiser. Using concept maps in teaching microbiology. 11(1):58–59, 2010. doi: 10.1128/jmbe.v1.i2.142. URL https://journals.asm.org/doi/10.1128/jmbe.v1.i2.142. Publisher: American Society for Microbiology.

Yunus Karakuyu. The effect of concept mapping on attitude and achievement in a physics course. 5(6):724–737, 2010. URL https://academicjournals.org/journal/IJPS/article-full-text-pdf/536E2A726759.pdf.

Annette Karmiloff-Smith. Beyond modularity: Innate constraints and developmental change. In *The epigenesis of mind: Essays on biology and cognition*, The Jean Piaget Symposium series, pages 171–197. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1991. ISBN 978-0-8058-0438-6.

Annette Karmiloff-Smith. *Beyond Modularity: A Developmental Perspective on Cognitive Science*. The MIT Press, September 1995. ISBN 9780262611145.

Maria Keet and Alessandro Artale. Representing and reasoning over a taxonomy of part–whole relations. *Applied Ontology*, 3(1):91–110, 2008.

Fred N. Kerlinger. *Foundations of behavioural research*. Holt, Rinehart and Winston, 3rd edn. edition, 1986.

Meena Kharatmal. Concept mapping for eliciting students' understanding of science. *Indian Educational Review*, 45(2):31–43, 2009. URL https://okeanos.files.wordpress.com/2009/11/concept-maps-meena-hbcse-ier-july-2009.pdf.

Meena Kharatmal. A proposal to extend concept mapping to concept lattices for representing biology. In *International Summer School on Methodology of Task Design – How to Construct Exercises for Learning,*, Dresden, Germany, 2014. Technische Universität. URL https://okeanos.files.wordpress.com/2014/12/summerschool-short-paper.pdf.

Meena Kharatmal. Field Property for Modeling Biological Process. In *Australian Systemic Functional Linguistics Association Conference 2023: Celebrating Semiosis*, 2023a. URL https://okeanos.files.wordpress.com/2023/12/asfla-2023-field-property.pdf.

Meena Kharatmal. Analysis of undergraduate biology students' multimodal depiction of experimental design using SFL framework. In *Australian Systemic Functional Linguistics Association Conference 2023: Celebrating Semiosis*, Australia, 2023b. URL https://okeanos.files.wordpress.com/2023/12/asfla-2023-multimodal-analysis.pdf.

Meena Kharatmal. Using SFL framework for nominalization of processes in biology textbooks. In *49th International Systemic Functional Congress*, 2024.

Meena Kharatmal and G. Nagarjuna. Understanding science through knowledge organizers. In Jayashree Ramadas and Sugra Chunawala, editors, *First International Conference to Review Research on Science, TEchnology and Mathematics Education*, volume 1, Mumbai, India, 2006a. HBCSE. URL https://okeanos.files.wordpress.com/2008/08/episteme-11.pdf.

Meena Kharatmal and G. Nagarjuna. A proposal to refine concept mapping for effective science learning. In A. J. Canas and J. D. Novak, editors, *Concept Maps: Theory, Methodology, Technology. Proceedings of the Second International Conference on Concept Mapping*, San Jose, Costa Rica, 2006b. URL https://okeanos.files.wordpress.com/2008/08/cmc2006-p151-2.pdf.

Meena Kharatmal and G. Nagarjuna. An alternative proposal for eliciting and assessing students' knowledge structure. In Chitra Natarajan and Beena Choksi, editors, *epiSTEME 2 – International Conference to Review Research Trends in Science, Technology and Mathematics Education*, Mumbai, India, 2007. MacMillan Publishers Ltd. URL https://okeanos.files.wordpress.com/2008/08/episteme-2.pdf.

Meena Kharatmal and G. Nagarjuna. 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, and J. D. Novak, editors, *Concept Mapping: Connecting Educators. Proceedings of the Third International Conference on Concept Mapping*, Tallinn, Estonia & Helsinki, Finland, 2008. URL https://okeanos.files.wordpress.com/2008/09/exploring-roots-of-rigor.pdf.

Meena Kharatmal and G. Nagarjuna. Refined concept maps for science education: A feasibility study. In K. Subramaniam and Anwesh Majumdar, editors, *Third International Conference to Review Research on Science, TEchnology and Mathematics Education*, volume 3, pages 76–80, Mumbai, India, 2009. HBCSE. URL https://okeanos.files.wordpress.com/2008/04/rcm-feasitibility-epi-3-12-meena-gn.pdf.

Meena Kharatmal and G. Nagarjuna. Introducing Rigor in Concept Maps. In Madalina Croitoru, Sébastien Ferré, and Dickson Lukose, editors, *Conceptual Structures: From Information to Intelligence*, volume 6208 of *Lecture Notes in Computer Science*, pages 199–202. Springer Berlin / Heidelberg, 2010. ISBN 978-3-642-14196-6. URL https://link.springer.com/chapter/10.1007/978-3-642-14197-3_22.

Meena Kharatmal and G. Nagarjuna. An Analysis of Growth of Knowledge Based on Concepts and Predicates — A Preliminary Study. In S. Chunawala and M. Kharatmal, editors, *Fourth International Conference to Review Research on Science, TEchnology and Mathematics Education*, Mumbai, India, 2011. HBCSE, Mumbai. URL https://okeanos.files.wordpress.com/2008/04/analysis-growth-knowledge-concepts-predicates-epi4-paper.pdf.

Meena Kharatmal and G. Nagarjuna. Representing Change Using Concept Maps. In *Proceedings of epiSTEME 5 – International Conference to Review Research on Science, Technology and Mathematics Education*, pages 124–131, Mumbai, India, 2013. Cinnamonteal. URL https://okeanos.files.wordpress.com/2008/04/representing-change-using-concept-maps-epi5-paper.pdf.

Meena Kharatmal and G. Nagarjuna. Using Semantic Reference Set of Linking Words for Concept Mapping in Biology. In *Innovating with Concept Mapping.*, volume 635 of *Communications in Computer and Information Science*, pages 315–329. Springer, Germany, 2016. URL https://link.springer.com/chapter/10.1007/978-3-319-45501-3_25.

Meena Kharatmal and G. Nagarjuna. A Preliminary Study on Effect of Providing Linking Words in Concept Mapping for Representing Cell Biology Knowledge. *Asian Journal of Biology Education*, 16:9–14, 2024. URL https://doi.org/10.57443/ajbe.16.0_9.

Meena Kharatmal, Mayur Gaikwad, Aashutosh Mule, and Jaikishan Advani. Exploring college biology students' understanding of experimental design. 14:2–7, 2022. URL https://doi.org/10.57443/ajbe.14.0_2.

Ian M Kinchin. Concept mapping in biology. 34(2):61–68, 2000. URL https://www.tandfonline.com/doi/abs/10.1080/00219266.2000.9655687.

Ian M Kinchin. If concept mapping is so helpful to learning biology, why aren't we all doing it? 23(12):1257–1269, 2001a.

Ian M. Kinchin. Can a novice be viewed as an expert upside-down?. 83(303):91–95, 2001b. URL http://eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ641852.

Ian M Kinchin, David B Hay, and Alan Adams. How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development. 42(1):43–57, 2000.

Bogumil M. Konopka. Biomedical ontologies—A review. *Biocybernetics and Biomedical Engineering*, 35(2):75–86, January 2015. ISSN 0208-5216. doi: 10.1016/j.bbe.2014.06.002. URL https://www.sciencedirect.com/science/article/pii/S0208521614000503.

Marta M. Koć-Januchta, Konrad J. Schönborn, Lena AE Tibell, Vinay K. Chaudhri, and H. Craig Heller. Engaging with biology by asking questions: Investigating students' interaction and learning with an artificial intelligence-enriched textbook. *Journal of Educational Computing Research*, 58(6):1190–1224, 2020. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Rob Kremer. The Design of a Concept Mapping Environment for Knowledge Acquisition and Knowledge Representation. In *Proceedings of the 9th International Knowledge Acquisition Workshop*, Banff, Canada, 1995.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology.* Sage Publications Inc., California, 2000. URL https://us.sagepub.com/en-us/nam/content-analysis/book258450.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology.* Thousand Oaks. Sage Publications., CA, 2nd ed. edition, 2004.

Thomas S. Kuhn. *The structure of scientific revolutions.* University of Chicago Press, April 1970. ISBN 9780226458038.

Tobias Kuhn. *Controlled English for knowledge representation.* PhD Thesis, University of Zurich, 2010.

Tobias Kuhn. A survey and classification of controlled natural languages. *Computational linguistics*, 40(1):121–170, 2014.

Vinayak G. Kulkarni. Role of language in science education. In Peter Fensham, editor, *Development and dilemmas in science education*, pages 150–167. The Falmer Press, London, 1988.

Margaret D. LeCompte and Judith Preissle. *Ethnography and qualitative design in educational research.* Academic Press, San Diego, 2nd edition, 1993.

Sangah Lee, Min Zhong, Cerrone Foster, Miriam Segura-Totten, and Melissa McCartney. From Novice To Expert: An Assessment To Measure Strategies Students Implement While Learning To Read Primary Scientific Literature. *Journal of Microbiology & Biology Education*, 23(3):e00126–22, November 2022. doi: 10.1128/jmbe.00126-22. URL https://journals.asm.org/doi/10.1128/jmbe.00126-22. Publisher: American Society for Microbiology.

Jay L. Lemke. *Talking Science: Language, Learning, and Values.* Praeger, Norwood, N.J, jan 1990. ISBN 9780893915667.

Douglas Lenat and Ramanathan V. Guha. CYC: A Midterm Report. *AI Magazine*, 11(3):32–32, 1990. ISSN 2371-9621. doi: 10.1609/aimag.v11i3.842. URL https://www.aaai.org/ojs/index.php/aimagazine/article/view/842. Number: 3.

Yvonna S. Lincoln and Egon G. Guba. *Naturalistic Inquiry.* Sage, 1985. URL https://doi.org/10.1016/0147-1767(85)90062-8.

Tony Long and Martin Johnson. Rigour, reliability and validity in qualitative research. *Clinical Effectiveness in Nursing*, 4(1):30–37, March 2000. ISSN 1361-9004. doi: 10.1054/cein.2000.0106. URL https://www.sciencedirect.com/science/article/pii/S1361900400901067.

Jiaying Lu, Xiangjue Dong, and Carl Yang. Weakly Supervised Concept Map Generation through Task-Guided Graph Translation. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–13, 2023. ISSN 1558-2191. doi: 10.1109/TKDE.2023.3252588. URL `https://ieeexplore.ieee.org/document/10059210`. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Dennis Luchembe, Kaumba Chinyama, and Jack Jumbe. The effect of using concept mapping on student's attitude and achievement when learning the physics topic of circular and rotational motion. 5(4):10–29, 2014. URL `https://files.eric.ed.gov/fulltext/EJ1068159.pdf`.

Seweryna Luszczewska-Romahnowa. Ambiguity and the Language of Science. In Jerzy Pelc, editor, *Semiotics in Poland 1984–1969*, Synthese Library, pages 148–159. Springer Netherlands, Dordrecht, 1979. ISBN 9789400997776. doi: 10.1007/978-94-009-9777-6_17.

Juliana Machado and Marco Braga. Representational Redescription as a framework for model conceptualization in Science Education. *Ciência & Educação (Bauru)*, 25: 589–606, 2019. Publisher: SciELO Brasil.

Robert Mack and Jill Robinson. When novices elicit knowledge: Question asking in designing, evaluating, and learning to use software. In *The psychology of expertise: Cognitive reserach and empirical AI*. Lawrence Erlbaum Associates, New Jersey, 1991.

Lorenzo Magnani, Nancy J. Nersessian, and Paul Thagard. *Model-based reasoning in scientific discovery*. Springer, 1999.

Arthur B. Markman and Edward J. Wisniewski. Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(1):54–70, 1997. ISSN 1939-1285(Electronic),0278-7393(Print). doi: 10.1037/0278-7393.23.1.54.

Judith A. Markowitz, J. Terry Nutter, and M. W Evens. Beyond is-a and part-whole: More semantic network links. *Computers & Mathematics with Applications*, 23(6-9): 377–390, 1992a.

Judith A Markowitz, J. Terry Nutter, and Martha W Evens. Beyond is-a and part-whole: More semantic network links. *Computers & Mathematics with Applications*, 23(6-9): 377–390, 1992b.

Barbara L. Martin, Joel J. Mintzes, and Ileana E. Clavijo. Restructuring knowledge in biology: cognitive processes and metacognitive reflections. 22(3):303–323, 2000. ISSN 0950-0693. doi: 10.1080/095006900289895. URL `http://www.tandfonline.com/doi/abs/10.1080/095006900289895`.

James R. Martin and David Rose, editors. *Working with Discourse: Meaning Beyond the Clause*. Continuum, New York, 2003.

James R. Martin, Len Unsworth, and David Rose. Condensing meaning: Imagic aggregations in secondary school science. 2021.

Ernst Mayr. *Principles of systematic zoology*. McGraw-Hill New York, 1969.

Mayring. *Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution.* Free Open Book, 2014.

William F. McComas. Nature of Science. In William F. McComas, editor, *The Language of Science Education: An Expanded Glossary of Key Terms and Concepts in Science Teaching and Learning*, pages 67–68. SensePublishers, Rotterdam, 2014. ISBN 978-94-6209-497-0. doi: 10.1007/978-94-6209-497-0_59. URL https://doi.org/10.1007/978-94-6209-497-0_59.

Deborah McGuinness. Ontologies come of age. In *Spinning the Semantic Web*, 2003.

Susan McKenney and Thomas Reeves. *Conducting Educational Design Research.* Routledge, London, 2 edition, August 2018. ISBN 978-1-315-10564-2. doi: 10.4324/9781315105642.

Carolyn B. Mervis and Eleanor Rosch. Categorization of Natural Objects. *Annual Review of Psychology*, 32(1):89, 1981. ISSN 0066-4308. URL https://www.academia.edu/20446775/Categorization_of_Natural_Objects.

George Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

Norma Miller and Alberto Canas. Effect of the nature of the focus question on presence of dynamic propositions in a concept map. In Alberto J. Canas, Priit. Reiska, Mauri Ahlberg, and Joesph D. Novak, editors, *Concept Mapping: Connecting Educators. Proceedings of the Third International Conference on Concept Mapping, Tallinn, Estonia & Helsinki, Finland*, Tallinn, Estonia & Helsinki, Finland, 2008.

Marvin Minsky. A Framework for Representing Knowledge. Technical Report Memo 306, MIT-AI Laboratory, USA, 1974.

Joel Mintzes and Heather J. Quinn. Knowledge restructuring in biology: Testing a punctuated model of conceptual change. 5(2):281–306, 2007. ISSN 1571-0068. URL http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ924413.

Joel Mintzes, James Wandersee, and Joseph Novak. *Teaching Science for Understanding: A Human Constructivist View.* Academic Press, 1998. ISBN 978-0-12-498361-8.

Joel Mintzes, James H. Wandersee, and Joseph D. Novak, editors. *Assessing Science Understanding: A Human Constructivist View.* Academic Press, 1 edition, 2000. ISBN 0-12-498365-0.

Joel J. Mintzes, James H. Wandersee, and Joseph D. Novak. Meaningful learning in science: The human constructivist perspective. In D. P. Gary, editor, *Handbook of academic learning: Construction of knowledge*, pages 405–447. Academic Press, USA, 1997.

Joel J. Mintzes, James H. Wandersee, and Joseph D. Novak. Assessing understanding in biology. 35(3):118–124, 2001. ISSN 0021-9266. doi: 10.1080/00219266.2001.9655759. URL http://www.tandfonline.com/doi/abs/10.1080/00219266.2001.9655759.

Elliot G. Mishler. Validation in inquiry-guided research: The role of exemplars in narrative studies. In Barbara M. Brizuela, Julie. P. Stewart, Romina G. Carrillo, and Jennifer G. Berger, editors, *Acts of inquiry in qualitative research*, pages 119–146. Harvard Educational Review, Cambridge, MA, 2000.

Jennifer L. Momsen, Tammy M. Long, Sara A. Wyse, and Diane Ebert-May. Just the facts? introductory undergraduate biology courses focus on low-level cognitive skills. 9 (4):435–440, 2010. doi: 10.1187/cbe.10-01-0001. URL `https://www.lifescied.org/doi/10.1187/cbe.10-01-0001`. Publisher: American Society for Cell Biology (lse).

Antonella Gabriella Montanaro, Filippo Sciarrone, and Marco Temperini. A Deep Learning Approach to Concept Maps Similarity. In *2022 26th International Conference Information Visualisation (IV)*, pages 239–244, July 2022. doi: 10.1109/IV56949.2022.00048. URL `https://link.springer.com/chapter/10.1007/978-3-031-33023-0_29`. ISSN: 2375-0138.

Sandy Muspratt and Peter Freebody. Understanding the Disciplines of Science: Analysing the Language of Science Textbooks. pages 33–59. 2013. ISBN 978-94-007-4167-6. doi: 10.1007/978-94-007-4168-3_3.

Ernest Nagel. *The Structure of Science: Problems in the Logic of Scientific Explanation*. Hackett, 1979. ISBN 9780915144716.

NCERT. NCERT, 1961. URL `https://ncert.nic.in/`.

NCERT. *National Curriculum Framework 2005*. NCERT, New Delhi, 2005.

NCERT. *Biology (Class XI Textbook)*. NCERT, New Delhi, 2007a.

NCERT. *Science (Class VIII Textbook)*. NCERT, NewDelhi, 2007b.

NCERT. *Science (Class IX Textbook)*. NCERT, New Delhi, India, 2007c.

NCERT. *Biology (Textbook for Class XII)*. NCERT, New Delhi, 2007d.

Nancy J Nersessian. Conceptual change. In William Bechtel and George Graham, editors, *A companion to cognitive science*, pages 155–166. Blackwell, 1998.

Mansoor Niaz and Arelys Maza. *Nature of Science in General Chemistry Textbooks*, volume 2 of *SpringerBriefs in Education*. Springer Netherlands, Dordrecht, 2011. ISBN 978-94-007-1919-4 978-94-007-1920-0. doi: 10.1007/978-94-007-1920-0. URL `https://link.springer.com/10.1007/978-94-007-1920-0`.

Joeseph Novak and Bob Gowin. *Learning How to Learn*. Cambridge University Press, USA, 1984.

Joseph D. Novak. Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching*, 27(10):937–949, 1990. ISSN 1098-2736. doi: 10.1002/tea.3660271003.

Joseph D. Novak. *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*. Taylor & Francis, second edition, 2010.

OBOFoundry. The Open Biological and Biomedical Ontologies, 2020. URL `http://www.obofoundry.org/docs/OperationsCommittee.html`.

Angela O'Donnell, Donald Dansereau, and Richard Hall. Knowledge maps as scaffolds for cognitive processing. 14:71–86, 2002. URL `10.1023/A:1013132527007`.

Andrew Olney, Whitney L. Cade, and Claire Williams. Generating Concept Map Exercises from Textbooks. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications, BEA 2011 at the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL-HLT*, pages 111–119., 2011.

Edison Ong, Zuoshuang Xiang, Bin Zhao, Yue Liu, Yu Lin, Jie Zheng, Chris Mungall, Mélanie Courtot, Alan Ruttenberg, and Yongqun He. Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic Acids Research*, 45(D1):D347–D352, January 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw918. URL https://academic.oup.com/nar/article/45/D1/D347/2770665. Publisher: Oxford Academic.

Relation Ontology. The OBO Relation Ontology, 2020. URL http://www.obofoundry.org/rodocs/OperationsCommittee.html.

Ontology-Summit. Ontology summit 2024, 2023a. URL https://ontologforum.com/index.php/OntologySummit2024.

Ontology-Summit. Ontology summit 2024, 2023b. URL https://groups.google.com/g/ontolog-forum/c/39aOYNyuBIQ.

Ontotext. Introduction to the semantic web — GraphDB SE 9.1 documentation, 1999.

Kathleen M. Palmquist, Michael E. Carley and Thomas A. Dale. Two applications of automated text analysis: Analyzing literary and non-literary texts. In Carl Roberts, editor, *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Tanscripts*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1997.

Michael Q. Patton. *Qualitative Research and Evaluation Methods*. Thousand Oaks, CA: Sage Publications, 2002a.

Michael Q. Patton. Two Decades of Developments in Qualitative Inquiry: A Personal, Experiential Perspective. *Qualitative Social Work*, 1(3):261–283, September 2002b. ISSN 1473-3250. doi: 10.1177/1473325002001003636. URL https://doi.org/10.1177/1473325002001003636. Publisher: SAGE Publications.

N. Renee Pearsall, Jo El J. Skipper, and Joel J. Mintzes. Knowledge restructuring in the life sciences: A longitudinal study of conceptual change in biology. 81(2): 193–215, 1997. ISSN 1098-237X. doi: 10.1002/(SICI)1098-237X(199704)81:2$⟨$193::AID-SCE5$⟩$3.0.CO;2-A.

John Piaget and Eleanor Duckworth. Genetic epistemology. *American Behavioral Scientist*, 13(3):459–480, 1970. URL http://abs.sagepub.com/content/13/3/459.3.short.

Karen Pine and David Messer. The development of representations as children learn about balancing. *British Journal of Developmental Psychology*, 21:285–301, June 2003. doi: 10.1348/026151003765264093.

Nikolay Podkolodnyy and Olga Podkolodnaya. Ontologies in bioinformatics and systems biology. *Russian Journal of Genetics: Applied Research*, 6:749–758, 11 2016. doi: 10.1134/S2079059716070091.

Argenta M. Price, Candice J. Kim, Eric W. Burkholder, Amy V. Fritz, and Carl E.

Wieman. A Detailed Characterization of the Expert Problem-Solving Process in Science and Engineering: Guidance for Teaching and Assessment. *CBE—Life Sciences Education*, 20(3):ar43, September 2021. doi: 10.1187/cbe.20-12-0276. URL https://www.lifescied.org/doi/10.1187/cbe.20-12-0276. Publisher: American Society for Cell Biology (lse).

John Priss, Uta Old. Concept neighbourhoods in lexical databases. In L. Kwuida and B. Sertkaya, editors, *Formal Concept Analysis. Proceedings of the 8th International Conference, ICFCA 2010, Morocco.*, volume 5986 of *Lecture Notes in Computer Science*, pages 282–295. Springer, 2010. URL https://link.springer.com/chapter/10.1007/978-3-642-11928-6_20.

M. Ross Quillian. Semantic Memories. In Marvin Minsky, editor, *Semantic Information Processing*, pages 216–270. MIT Press, Cambridge, MA, 1968.

Willard Van Orman Quine. On What There Is. *Review of Metaphysics*, 2(5):21–38, 1948. doi: revmetaph19482115.

Peter Mark Roget. *The Original Roget's Thesaurus of English Words and Phrases*. Longmans, Green & Co. / Dell Publishing Co., New York, 1962.

Alexander Rosenberg. *The Structure of Biological Science*. Cambridge University Press, Cambridge, 1985. ISBN 9780521275613. doi: 10.1017/CBO9781139171724.

Cornelius Rosse and José L. V. Mejino. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36(6):478–500, December 2003. ISSN 1532-0464. doi: 10.1016/j.jbi.2003.11.007.

Daniel L. Rubin, Nigam H. Shah, and Natalya F. Noy. Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics*, 9(1):75–90, January 2008. ISSN 1477-4054. doi: 10.1093/bib/bbm059.

Maria Araceli Ruiz-Primo, Susan E. Schultz, Min Li, and Richard J. Shavelson. Comparison of the reliability and validity of scores from two concept-mapping techniques*. *Journal of Research in Science Teaching*, 38(2):260–278, 2001. ISSN 1098-2736. doi: 10.1002/1098-2736(200102)38:2$⟨$260::AID-TEA1005$⟩$3.0.CO;2-F.

Craig Schlenoff, Michael Gruninger, Florence Tissot, John Valois, Joshua Lubell, and Jonathan Lee. The process specification language (psl) overview and version 1.0 specification, 2000-02-01 00:02:00 2000. URL https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=821500.

Norbert Schmitt. Current perspectives on vocabulary teaching and learning. *International handbook of English language teaching*, pages 827–841, 2007.

Stefan Schulz, Philipp Daumke, Barry Smith, and Udo Hahn. How to distinguish parthood from location in bioontologies. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2005:669–673, 2005. ISSN 1942-597X.

Ulf Schwarz and Barry Smith. Ontological Relations. In Kathrine Munn and Barry Smith, editors, *Applied Ontology: An Introduction*, pages 219–234. De Gruyter, May 2013. URL https://doi.org/10.1515/9783110324860.219.

Rolf Schwitter. Controlled natural languages for knowledge representation. In *Coling 2010: Posters*, pages 1113–1121, 2010.

Clive Seale. Quality in Qualitative Research. *Qualitative Inquiry*, 5(4):465–478, December 1999. ISSN 1077-8004. doi: 10.1177/107780049900500402. URL https://doi.org/10.1177/107780049900500402. Publisher: SAGE Publications Inc.

Barry Smith and Christiane Fellbaum. Medical WordNet: A New Methodology for the Construction and Validation of Information Resources for Consumer Health. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 371–382, Geneva, Switzerland, August 2004. COLING. URL https://www.aclweb.org/anthology/C04-1054.

Barry Smith, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L Rector, and Cornelius Rosse. Relations in biomedical ontologies. *Genome biology*, 6(5):R46, 2005.

Barry Smith, Waclaw Kusnierczyk, Daniel Schober, and Werner Ceusters. Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain. In *Proceedings of KR-MED, CEUR, vol. 222*, pages 57–65. 2006. URL https://philarchive.org/rec/SMITAR.

Edward E. Smith and Douglas L. Medin. *Categories and Concepts:*. Harvard University Press, Cambridge, MA, 1981. ISBN 978-0-674-86627-0.

Larisa Soldatova. The overuse of is-a relation, April 2010. URL http://ontogenesis.knowledgeblog.org/699.

John Sowa. Concept mapping. San Fancisco, USA, 2006. URL http://www.jfsowa.com/talks/cmapping.pdf.

John F. Sowa. *Conceptual structures: information processing in mind and machine*. Addison-Wesley Publishing Company, USA, 1984.

John F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks / Cole, USA, 001 edition, 2003. ISBN 0534949657.

Caroline Stenbacka. Qualitative Research Requires Quality Concepts of Its Own. *Management Decision,*, 39:551–555, 2001.

Michael. Stubbs. *Educational Linguistics*. Blackwell, Oxford, 1986.

Dennis J. Taylor, Nigel P. O. Green, G. Wilfred Stout, and Roland Soper. *Biological Science*. Cambridge University Press, UK, 3 edition, 2003. ISBN 0521561787.

Paul Thagard. *Conceptual Revolutions*. Princeton University Press, 1992. ISBN 9780691024905.

Rodney B. Thiele and David F. Treagust. Analogies in chemistry textbooks. *International Journal of Science Education*, 17(6):783–795, 1995. Publisher: Taylor & Francis.

George Tsatsaronis, Iraklis Varlamis, and Michalis Vazirgiannis. Word Sense Disambiguation with Semantic Networks. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, Lecture Notes in Computer Science, pages 219–226, Berlin, Heidelberg, 2008. Springer. ISBN 978-3-540-87391-4. doi: 10.1007/978-3-540-87391-4_29.

Len Unsworth. Intermodal relations, mass and presence in school science explanation genres. In Michele Zappavigna and Shoshana Dreyfus, editors, *Discourses of Hope and Reconciliation: On J. R. Martin's Contribution to Systemic Functional Linguistics*, pages 131–151. Bloomsbury Publishing., 2020. URL `https://doi.org/10.5040/9781350116092.ch-007`.

Necla Dönmez Usta, Eser Ültay, and Neslihan Ültay. Reading the concept map of physics teacher candidates: A case of light. 31(1):14–21, 2014. URL `https://doi.org/10.33828/sei.v31.i1.2`.

Georg Henrik von Wright. *Norm and action: a logical enquiry*. Humanities, 1963.

Stella Vosniadou, editor. *International Handbook of Research on Conceptual Change*. Taylor & Francis, 2008. ISBN 9780805860443.

Chronoula Voutsina. A micro-developmental approach to studying young children's problem solving behavior in addition. *The Journal of Mathematical Behavior*, 31(3): 366–381, September 2012. ISSN 0732-3123. doi: 10.1016/j.jmathb.2012.03.002. URL `https://www.sciencedirect.com/science/article/pii/S073231231200017X`.

James H. Wandersee. Using concept mapping as a knowledge mapping tool. In *Mapping biology knowledge*, pages 127–142. Springer, 2002. URL `http://link.springer.com/content/pdf/10.1007/0-306-47225-2_9`.

Feng Wang and Michael Hannafin. Design-based research and technology-enhanced learning environments. Educational Technology Research and Development, 53(4), 5-23. *Educational Technology Research and Development*, 53:5–23, December 2005. doi: 10.1007/BF02504682.

Yi-Hsuan Wang. Design-based research on integrating learning technology tools into higher education classes to achieve active learning. *Computers & Education*, 156: 103935, October 2020. ISSN 0360-1315. doi: 10.1016/j.compedu.2020.103935. URL `https://www.sciencedirect.com/science/article/pii/S0360131520301330`.

Robert P. Weber. *Basic content analysis*. Sage, New Delhi, 1985. URL `https://us.sagepub.com/en-us/nam/basic-content-analysis/book3202`.

Morton E Winston, Roger Chaffin, and Douglas Herrmann. A taxonomy of part-whole relations. *Cognitive science*, 11(4):417–444, 1987.

Ludwig Wittgenstein. *Tractatus Logico-Philosophicus*. Kegan Paul, London, 1922. URL `http://www.gutenberg.org/ebooks/5740`.

Rachel Wong, NarayanKripa Sundararajan, Olusola Adesope, and Krista Nishida. Static and interactive concept maps for chemistry learning. 41(2):206–223, 2021. URL `https://www.tandfonline.com/doi/full/10.1080/01443410.2020.1761299`.

Joseph H. Woodger. The Axiomatic Method in Biology. *Annals of the Entomological Society of America*, 31(2):210, June 1938. ISSN 0013-8746. doi: 10.1093/aesa/31.2.210.

William A. Woods. *What's in a link: Foundations for semantic networks*. Readings in cognitive science: A perspective from psychology and artificial intelligence. Morgan

Kaufmann, San Mateo, CA, US, 1988. ISBN 978-1-55860-013-3. URL https://www.sciencedirect.com/science/article/abs/pii/B9780121085506500070. Pages: 125.

Yang Xiang and Sarath Chandra Janga. Building Integrated Ontological Knowledge Structures with Efficient Approximation Algorithms. 2015.

Che-Yu Yang. Word sense disambiguation using semantic relatedness measurement. *Journal of Zhejiang University-SCIENCE A*, 7(10):1609–1625, October 2006. ISSN 1862-1775. doi: 10.1631/jzus.2006.A1609. URL https://doi.org/10.1631/jzus.2006.A1609.

Jenna M. Zukswert, Megan K. Barker, and Lisa McDonnell. Identifying Troublesome Jargon in Biology: Discrepancies between Student Performance and Perceived Understanding. *CBE Life Sciences Education*, 18(1):ar6, 2019. ISSN 1931-7913. doi: 10.1187/cbe.17-07-0118.

Joost Zwarts and Peter Gärdenfors. Locative and Directional Prepositions in Conceptual Spaces: The Role of Polar Convexity. *Journal of Logic, Language and Information*, 25 (1):109–138, March 2016. ISSN 1572-9583. doi: 10.1007/s10849-015-9224-5.

Jan Zytkow. Scientific Modeling: A Multilevel Feedback Process. In Lorenzo Magnani, Nancy Nersessian, and Paul Thagard, editors, *Model-Based Reasoning in Scientific Discovery*. Springer, New York, 1999.

List of Appendices

Appendix A

Dimension	Linking Words	Ontology
Part-whole	Associated with	Subcellular Anatomy Ontology
	Composed of	Relations Ontology
	Consists of/part of	Cell Line Ontology
	Constitute/ Composition	NIF Dysfunction
	Exists	Fission Yeast Phenotype Ontology
	Is interrupted by/ Discontinuous	NIF Cell Ontology
	Is Perforated by	Fission Yeast Phenotype Ontology
	Lacks/ absent in	Foundational Model of Anatomy
	Present in/ Found in/ Observed in	Phenotypic Quality

Dimension	Linking Words	Ontology and URL
Class Inclusion	IS-A	Relations

Dimension	Linking Words	Ontology and URL
Spatial Inclusion	Aligned/ is lined	Relations Ontology
	Arranged	Plant Trait Ontology
	Attached to	Relations Ontology
	Between	Human Phenotype Ontology
	Bound by/ Bound to	Foundational Model of Anatomy
	Boundary of	Subcellular Anatomy Ontology
	Connect / interconnected	Relations Ontology
	Contained in	CellLine Ontology
	Continuous	Fission Yeast Phenotype Ontology
	Covered	Cephalopod Ontology
	Enclosed	Human Phenotype Ontology
	Envelopes	Gene Ontology
	Exit through/ Protruding out of	Phenotypic Quality
	Extends	Phenotypic Quality
	Separated from	NIF Cell
	Located between/in	Relations Ontology
	Occupies	Vaccine Ontology
	Occurs	Ontology of Biological Attributes
	Projects into	Vaccine Ontology
	Surrounded by	Cell Line Ontology
	Wound around	Cell Line Ontology

Dimension	Linking Words	Ontology and URL
Function	Carries/ Channel for	Neoru Behavior Ontology
	Developed from	Bilateria Anatomy
	Formed by	Foundational Model of Anatomy
	Has Function/ Helps/ Performs	Biological Collections Ontology
	Has Role/Plays Role	Biological Collections Ontology
	Is site for	Cell Line Ontology
	Produces/ Synthesized from	Cell Ontology
	Reproduced by/ in	VIVO ISF
	Results in	Semanticscience Integrated Ontology

Dimension	Linking Words	Ontology and URL
Attributes	Has color	NIF Dysfunction
	Has age	Ontobee -DINTO
	Has density	NIF Cell
	Has diameter	Phenotypic Quality
	Has length	Phenotypic Quality
	Has number/ count	NIF Dysfunction
	Has property	Physico-chemical Methods and Properties
	Has shape	Cell Ontology
	Has size	Cell Ontology

Appendix B

Glossary of Relations

We provide a list of well-defined relations that are widely used in the thesis work. The major sources are from:

1. NCBO – `https://bioportal.bioontology.org/ontologies`
2. OBO Foundry – `http://www.obofoundry.org/`
3. Ontobee – `http://www.ontobee.org/`
4. Aber-OWL – `http://aber-owl.net/ontology/GO`
5. OLS – `http://www.ebi.ac.uk/ols/beta/ontologies/go`
6. AmiGO – `http://amigo.geneontology.org/`

The following are some of the critical relations referred from Smith et al. (2005).

is a =def. For continuants: C is_a C' if and only if: given any c that instantiates C at a time t, c instantiates C' at t. For processes: P is_a P' if and only if: that given any p that instantiates P, then p instantiates P.'

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

contained in =def. C contained_in C' if and only if: given any instance 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' at t, and it is not the case that c *overlaps* c' at t. (c' is a conduit or cavity).

is instance of =def. An object is an instance of a class if it is a member of the set denoted by that class. One would normally state the fact that individual i is an instance of class C with the relational form (C i), but it is equivalent to say (INSTANCE-OF i C)

composed_of=def. Component parts of anatomy of tissue made up of certain cells or other body area/system or tissue types.

composition=def. A single physical entity inhering in an bearer by virtue of the bearer's quantities or relative ratios of subparts.

enveloped by: [definition]. X enveloped_by y if and only if (1) y is a part of x, and (2) x is surrounded by y

Surrounded by = [definition]. X surrounded_by y if and only if (1) x is adjacent to y and for every region r that is adjacent to x, r overlaps y (2) the shared boundary between x and y occupies the majority of the outermost boundary of x

c located_in r at t - [definition]. a primitive relation between a continuant instance, a spatial region which it occupies, and a time

proposed: c located_on r at t - [definition]. a primitive relation between a continuant instance, a spatial region on which it occupies, and a time

The "location" relation of the spatial inclusion relation has been also formally defined as when each physical object or anatomical entity (c, d) is associated with a spatial region (r)at a time (t) as: located-in(c, d,t) = [definition]. part-of(r(c,t), r(d,t),t) (Schulz et al., 2005)

immersed in: [definition] a relation between a (physical) entity and a fluid substance in which the entity is wholely or substantially surrounded by the substance.

completely_varies_in: [definition]. A relation between two sequences or features that are considered variant with each other along their entire extents. (source: http://purl.obolibrary.org/obo/GENO_0000784)

proposed definition: varies_in: [definition] When X varies_in Y, it implies instances of X can occur in variable instances of Y.

c contained_in c 1 at t = [definition] c located_in c 1 at t and not c overlap c 1 at t. (Smith et al., 2005)

contained-in(c, d,t) = [definition]. located-in(c, d,t) ¬part-of(c, d,t)

Glossary of Attributes

We provide a list of well-defined attributes that are widely used in the thesis work. The major sources are from:

1. Ontobee – http://www.ontobee.org/

    length = [definition]. A 1-D extent quality which is equal to the distance between two points

    diameter = [definition]. A length quality which is equal to the length of any straight line segment that passes through the center of the circle and whose endpoints are on circular boundary.

    size = [definition]. A morphological quality inhering in a bearer by virtue of the bearer's physical magnitude.

    shape = [definition]. A morphological quality inhering in a bearer by virtue of the bearer's ratios of distances between its features (points, edges, surfaces and also holes etc.).

    color = [definition]. A composite chromatic quality composed of hue, saturation and intensity parts.

    temperature = [definition]. A physical quality of the thermal energy of a system.

    density = [definition]. The amount of something per unit size.

2. Library of Generic Concepts - https://www.ai.sri.com/~halo/public/clib/20130328/components/specs/slotdictionary.html#Entity-Property-Value

    specific surface area = [definition]. Property for a spatial entity to measure the total surface area per unit of its mass or volume.

    fluidity = [definition]. Property indicating relative ease of molecules associated with an object to freely move past one another.

surface tension = [definition]. The property 'Surface tension' of a liquid substance refers to the ability of its surface to resist an external force.

luminous flux = [definition]. The rate of flow of light energy.

magnification = [definition]. The amplification of the view of a physical object by magnifying an area for greater detail.

Glossary of Processes, States

We provide a list of well-defined process specifications that are widely used in the thesis work.

1. Process Specification Language - `https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir6459.pdf`

state = [definition]. The state relation holds between a fluent and an activity-occurrence, if the fluent holds before the activity-occurrence.

post-state = [definition]. The effect relation holds between a fluent and an activity if the fluent holds after the activity-occurrence.

during = [definition]. occurrence1 and occurrence2 are two activity- occurrences that are subactivity occurrences of occurrence3 and that the beginning and ending timepoints of occurrence1 are between the beginning and ending timepoints of occurrence2.

overlaps = [definition]. occurrence1 and occurrence2 are two activity occurrences that are subactivity occurrences of occurrence3 and that the beginning timepoint of occurrence1 is between the beginning and ending timepoints of occurrence2, and the ending timepoint of occurrence2 is between the beginning and ending timepoint of occurrence1.

changes = [definition]. This relation is used to capture the effect of an activity-occurrence on the properties of the world. An activity-occurrence changes a fluent if either the fluent held before the activity-occurrence and does not hold after it, or the fluent did not hold before the activity-occurrence but holds after it.

Web Resources of Concept Graphs

`https://www.ai.sri.com/~halo/public/exported-kb/graphs/page.html`

Appendix C

Semantic Categories

The list is not exhaustive even for a given domain. This serves the limited purpose of demonstrating the feasibility of the relational content analysis suggested. If the idea finds any value, one may extend, or modify the set of rules and the corresponding codes.

Semantic categories:

Object Properties:

C1 Meronymic inclusion [has part, composed of]

C2 Class inclusion [is a]

C3 Class member [instance of]

C4 Spatial inclusion [covered by, located in]

C5 Function [role/has function]

C6 Process [changes into, develops from, becomes, develops, moves towards, before, results into]

Data Properties:

C7 Geometric Form [shape, appearance]

C8 Measurable [size, length, diameter, width]

C9 Quality [color]

Re-representation Rules

Re-representation *rules*:

R1 Pass those statements that already use the object and data properties available in the Reference Set appropriately.

R2 Eliminate ambiguity of common verbs, <u>is</u>, <u>are</u>, restricting their lone usage to an unambiguous category of semantic relation for subtyping or use them only in combination with the object and data properties available in the Reference Set.

R3 Eliminate ambiguity of common verbs, <u>has</u>, and <u>have</u>, restricting their lone usage. Based on the intended meaning, use verbs <u>has</u>, and <u>have</u> always in combination with data property, or object property as appropriate.

R4 Eliminate the lone usage of prepositions as semantic relations between subject and object, and use them only in combination with the object and data properties available in the Reference Set and to create complex expressions between class terms.

R5 Rewrite the adjectives in the form of data properties or its inverse names such as <u>size</u>, <u>color</u>, <u>shape</u>, <u>length</u>, <u>form</u>, <u>density</u>, <u>pH</u>, <u>viscosity</u>, <u>solubility</u>, etc.

R6 Rewrite the verbs that represent processes/events into nouns, by using object properties such as <u>has role/has function</u>, <u>has sub-process</u>, <u>is site of</u>, etc.

R7 Rewrite the verbs in the form of object properties or their inverse names, after nominalization when necessary.

R8 Represent the change in the events as prior-state and post-state following temporal order.

R9 Add prepositions to create condensed expressions.

R10 Relating statements through logical relations such as because, therefore, and, hence, or, etc.

Illustrations

We illustrate the relational content analysis starting with simple to complex examples. We provide a list of re-representation rules [Rn], and a list of categories of semantic relations [Cn] which are used for coding while performing content analysis. A compiled list of re-representation rules and categories are presented above. We follow the notation of numbering the source text as (n) and its re-represented form as ($n\prime$). In the re-represented form, we code the underlined verb as $[Rn, Cn]$ to indicate re-representation rules and coding of semantic categories applied respectively. The illustrations are organized as text passage followed by re-represented passages assigned with the codings. Each case of illustration is followed by discussion and justification for the re-representation and codings.

1. A phospholipid is an amphipathic molecule.

re-written as:

(1′) A phospholipid is an [R1, C2] amphipathic molecule.

In sentence (1) the verb is is used to express semantic relation class inclusion, which is one of the core relations available in the Reference Set. The relation is expressed between two classes, where the members of the former class are included in the latter class. A formal definition of object property is is as follows:

> is_a =def. For continuants: C is_a C' if and only if: given any c that instantiates C at a time t, c instantiates C' at t. For processes: P is_a P' if and only if: that given any p that instantiates P, then p instantiates P' (Smith et al., 2005).

It may be noted that is is defined for both continuants (for a class of objects) and occurents (for a class of events/processes). Since the intended meaning matches with the verb as defined in the Reference Set, the sentence does not require re-representation and therefore it is coded as [R1].

2. Each mitochondrion is a double membrane-bound structure with the outer membrane and the inner membrane.

re-written as:

(2′) Each mitochondrion is an instance of [R2, C3] a double membrane-bound structure, which consists of [R4, C1] the outer membrane and the inner membrane.

In the sentence (2), the same verb term is is used as an instance to express class membership relation between 'each mitochondrion' and 'membrane bound structure'. Since, we restrict the use of one verb for one meaning to eliminate ambiguous usage, we re-represent it with a well defined object property is instance of with the code [R2, C3]. The relation is defined as follows:

> is instance of =def. An object is an instance of a class if it is member of the set denoted by that class. One would normally state the fact that individual i is an instance of class C with the relational form (C i), but it is equivalent to say (INSTANCE-OF i C) (Smith et al., 2005).

Further, the preposition with is used for expressing the part and whole relationship. We restrict the use of prepositions along with the object or data property. Therefore we eliminate with and replace it with consists of, which is coded as [R4] and as [C1] because it expresses meronymy.

The object property consists of/has part or or its inverse relation part of, is defined as:

> 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. The "part of" definition asserts that c is an instance of C exists in at a given time t, there exists c1 an instance of C1 at the same time, and c1 is part of c. It also asserts that c does not exist except as part of c1 (Smith et al., 2005).

3. Other types of membrane lipids are also amphipathic.
4. The leucoplasts are the colorless plastids of varied shapes and sizes with stored nutrients.
5. Typically [mitochondrion] is sausage-shaped or cylindrical having a diameter of 0.2 - 1.0 micrometer (average 0.5 micrometer) and length 1.0 - 4.1 micrometer.

re-written as:

(3′) Other types of membrane lipids also have amphipathic property [R2, C9].

(4′) The leucoplasts are [R1, C2] the colorless plastids. They have various shapes and sizes [R3, C7, C8] and contain [R4, C4] stored nutrients.

(5′) Mitochondrion has sausage or cylindrical shape, [R2, C7] has a diameter of [R3, C8] of 0.2 - 1.0 micrometer (average 0.5 micronmeter) and has length of [R3, C8] 1.0 - 4.1 micrometer.

In the sentences from (3) to (5), the verbs is/are are used in four different senses: class inclusion, data property, spatial inclusion, and class membership. Researchers have reported the use of the verb is for genus-subsumption, determinable-subsumption, specification, specialization (Johansson, 2013). Ambiguity arises when is is used for expressing meronymic inclusion, spatial inclusion and attribution (data property) (Kharatmal and Nagarjuna, 2010). Therefore, it calls for re-representation.

In sentence (3), the verb are is being used for the intended meaning of data property. For all data properties, we consistently use 'has property/have properties'. This requires re-representation of the verb are to have property. This is coded as [R2, C9].

Sentence (4) is a complex sentence, the subject "leucoplasts" is predicated for class inclusion, shape, size as well as what it contains. The same sentence is paraphrased with the common subject i.e. "leucoplast", and can give rise to at least three simple statements/propositions – (i) leucoplasts are colorless plastids; (ii) leucoplasts are of varied shapes and sizes; (iii) leucoplasts are with stored nutrients. In statement (i), the intended meaning is class inclusion. Therefore it does not require re-representation. In (ii), the preposition of is used to express a data property. While in (iii) the preposition with is used to express the containment relation. It may be noted that the same preposition with in example (2) was used to express meronymy while in this case it is used to express spatial inclusion relation resulting in ambiguity. This strengthens our argument to eliminate the use of prepositions and replace these with appropriate semantic relations to make the meaning explicit and unambiguous.

In sentence (5), the intended meaning is about describing the shape, diameter, and length as data properties. The verb is is replaced with has shape and the verb having is replaced with has a diameter of, has a length of and coded accordingly.

6. Plastids also have their own DNA and ribosomes.
7. The membrane of the erythrocyte has approximately 52 percent protein and 40 percent lipids.
8. Mitochondria have two membrane coverings instead of just one.

re-written as:

(6′) Plastids also consist of [R3, C1] their own DNA and ribosomes.

(7′) The membrane of the erythrocyte is composed of [R3, C1] approximately 52 per cent protein and 40 per cent lipids.

(8′) Mitochondria is covered by [R3, C4] two membranes instead of just one.

In the sentences from (6) to (8), the same verb has/have is used to for expressing substantially different semantic relations: consists of in (6), composed of in (7) and covered by in (8). Though both (6) and (7) are apparently similar i.e. belonging to meronymic inclusion, (6) asserts one being a *part* of another, while (7) asserts one is a *stuff* of another (Winston et al., 1987). In (8) the intended meaning is a type of spatial inclusion and therefore covered by is used. Thus in (6), (7), and (8), we note that the verb have is used to express substantially different meanings, and also in (5). By using only one verb have we cannot capture the wide variety of conceptual spaces (predicate spaces) required for modeling scientific phenomena. Therefore, (6), (7), and (8) becomes (6′), (7′), and (8′) respectively after re-writing.

The definition of object properties "composed of", and "composition of" is presented below:

> composed of=def. Component parts of anatomy of tissue made up of certain cells or other body area/system or tissue types. composition=def. A single physical entity inhering in an bearer by virtue of the bearer's quantities or relative ratios of subparts. We can apply this definition in the above example: as A single physical

entity (membrane of erythrocyte) inhering (comprising) ratios of subparts (proteins and lipids). (Ong et al., 2017)

9. A peripheral proteins lie on the surface of membrane while the integral proteins are partially or totally buried in the membrane.
10. The nonpolar tail of saturated hydrocarbons is protected from the aqueous environment.

re-written as:

(9′) A peripheral proteins located on [R4, C4] surface of membrane; Integral proteins are partially or totally embedded in [R4, C4] the membrane.

(10′) The non polar tail of saturated hydrocarbon; Non polar tail has role in [R6, C5] protection from the aqueous environment.

The sentence (9) paraphrased as: (i) "Peripheral proteins lie on surface of membrane"; (ii) "Integral proteins are partially or totally buried in the membrane". The statements express the place (location) of the objects within other objects. According to (Winston et al., 1987), these are spatial inclusion relations, wherein "object is located or surrounded but not a part of the object that it locates or surrounds it". The objects (entities) that occupy in the spatial regions are re-written with relations "located in", "contained in", and "adjacent to" (Smith et al., 2005). Although the verbs lie on and located on are synonymous with each other, for consistency we will continue to use located on, as it is part of the Reference Set.

The subject of the sentence (10) is a complex expression referring to the structure that has a role. To highlight this, the re-written sentence uses the semantic relation has a role in. The verb protected is nominalized into 'protection'.

In all the above examples, we largely considered sentences that describe structures. Biological text largely includes a description of structure and function. Biology text also has passages that describe processes. We shall perform a relational content analysis of process-centric sentences.

11. Cell membranes separates contents of cells from their external environments, controlling exchange of materials such as nutrients and waste products between the two. They also enable separate compartments to be formed inside cells in which specialized metabolic processes such as photosynthesis and aerobic respiration can take place. Chemical reactions, such as light reactions of photosynthesis in chloroplasts, sometimes take place on the surface of the membranes themselves.

The passage is re-written as:

(11′) Cell membranes have a role in [R6, R7, C5] the separation of contents of a cell from their external environments, the formation of separate compartments inside a cell, the control of material exchange, and [R10] nutrients and waste products between [R1], the two. Special metabolic processes includes [R2, C2], photosynthesis and aerobic respiration, which occur in [R1, C4] the separate compartments of cells. Chemical reactions includes

[R2, C2] light reactions, occur in [R1, C4] chloroplasts during photosynthesis, which occur on [R1, C4] surface of membranes.

Passage (11), describes the function of the cell membrane. Therefore, we replace the verb forms by nominalization (Halliday, 2006), where separates becomes "separation of content", forming becomes "formation of", and controlling becomes "control of" by bringing them to the subject place in the sentence. These nominalized process terms occur as class terms in ontologies (Ong et al., 2017). In the nominalized statement, the verbs are replaced by a well defined predicate "has role/has function" (Arp and Smith, 2008b) available from the Reference Set. The re-written statements depict the processes at object place (although these can be at either object or subject place) Further, with just an additional re-arrangement (Juznic, 2012), these process terms can be brought to focus in subject place, using the inverse name of the object property, while preserving the meaning. For example, "The separation of contents of cell is the role of a cell membrane".

The term such as has the sense of subtyping, therefore we replace it with defined object property includes (an inverse name of object property is). The verbs takes place in depict the sense of site or location. Although it's expression is correct, we use its synonym occur in as specified in the Reference Set. One may also note that the prepositions in and on indicate locative prepositions (Zwarts and Gärdenfors, 2016) and following these have been rewritten in combination with explicit predicate terms. The conjunction and is used to combine the two subjects. The relation between is a well defined three place predicate with three arguments (Sowa, 1984), and therefore remains unchanged.

Scientists approach phenomena/processes not merely by naming the terms but modeling by identifying the various parameters/variables, where the change is modeled in terms of change in state.

Consider the following passage:

12. The beginning of prophase is indicated by the appearance of the chromosomes as thin threads inside the nucleus. In fact, the word "mitosis" (Gr., mitos, thread) is an expression of this phenomenon, which becomes more evident as the chromosomes start to condense. The condensation occurs by a process of folding of the chromatin fibers. At the same time, the cell becomes spheroid, more refractile, and viscous. (De Robertis and De Robertis, 1985, p. 420).

The passage is re-written as:

(12′) Chromatin condensation begins in [R8, C6] prophase and ends in [R8, C6] metaphase. During condensation the unfolded chromatin becomes [R8, C6, C7] folded. The thin threads of chromosomes becomes condensed chromosomes. During the process, the cell becomes spheroid, more refractile and viscous. The term "mitosis" refers to [R1] the thread formation, which is derived from [R1] "mitos" meaning thread.

The passage refers to several state changes. To retain the temporal order, we modeled the process as prior state and post state as a requirement of process modeling elaborated in the previous section. We use becomes as a semantic relation between prior state and post state. "The unfolded chromatin becomes folded" or "thin threads becomes condensed, thick chromosomes". Similarly, refers to is a defined object property for naming. Two or more class terms or adjectives can be combined with the usage of prepositions to create condensed expressions. This passage represents processes as a state change model.

To summarize, we presented an exploratory study on relational content analysis of science text focusing on object properties and data properties. The emphasis is on rewriting the sentences to arrive at an explicit form. This re-statement of a sentence is categorized as *explicitly stated knowledge* in KR (Chaudhri et al., 2011). As part of the re-representation of relations, our approach has been towards resolving ambiguity, making implicit meaning explicit, making distinctions between object and data property, nominalization of verbs for processes, and representing processes in terms of state change. We followed the re-representation rules and assigned codes.