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

Meena Kharatmal

Synopsis

Submitted in partial fulfillment of the requirements for the Degree of
Doctor of Philosophy in the Homi Bhabha Centre for Science Education,
Tata Institute of Fundamental Research, Mumbai, India

July 2023

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 work, I 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 six chapters, starting with an introduction and motivation in chapter 1. I 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 I 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, I 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, I 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, I 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. With this, I frame research questions asking about the significance of this method in growth of knowledge, to test its feasibility, to provide indicator for developing expertise, all of which are relevant for science education research. I discuss these studies, data analysis, and findings in chapter 5. In the final chapter, I 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.

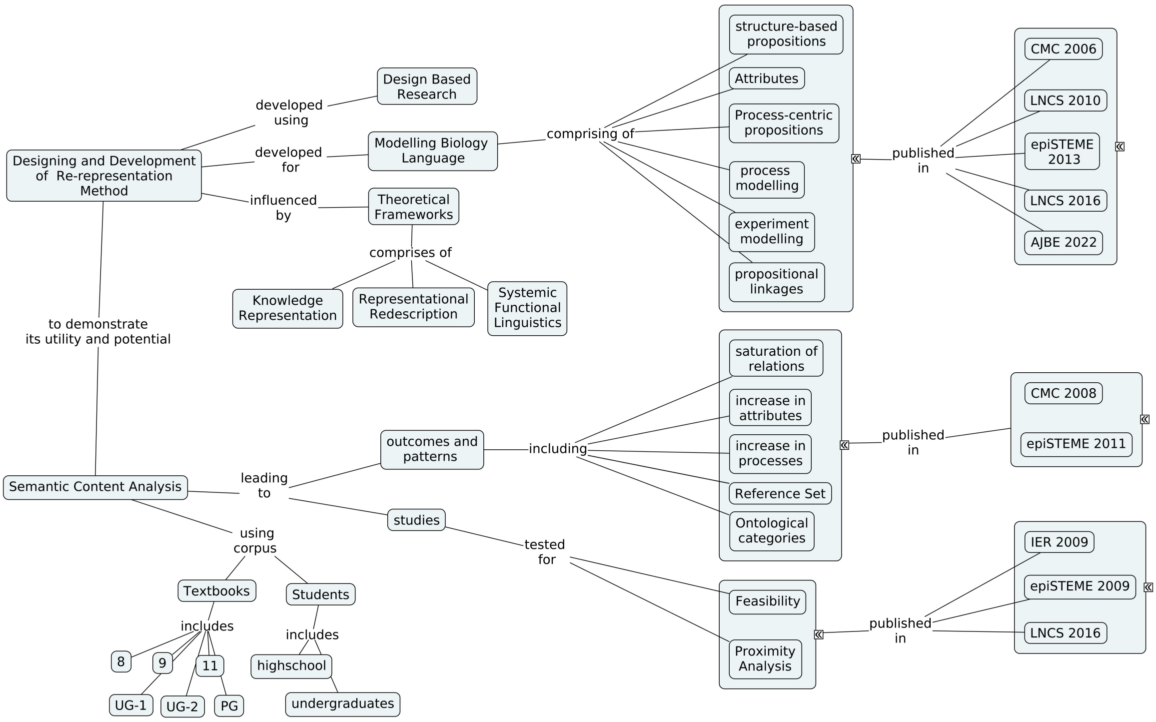

Figure 1: A graphical representation of the research model. (To read from top to bottom and left to right).

Peer-reviewed publications arising from the work in this thesis

1. 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

2. 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. Springer. DOI: 10.1007/978-3-319-45501-3_25

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

4. Meena Kharatmal (2009): Concept Mapping for Eliciting Students' Understanding of Science. Indian Educational Review, 45(2), pp.31-43.

Synopsis

This thesis is about the language of science, specifically biology, from the perspective of relational content analysis of biology textbooks and students' representations. We identified gaps and problems in the nature and usage of predicates used for connecting concepts and objects in sentences. To address these gaps, a method of re-representation was designed and developed with iterative refinements focusing on the semantics of predicates, inspired by semantic network-based knowledge representation, concept mapping, representational redescription, and systemic functional linguistics. The method was validated by using well-defined predicates from a published open biomedical repository. The method was used for modeling biology language in terms of predicates used for describing structures, processes, variable properties, experimental procedures, and explanations. This resulted in curating a Reference Set of predicates (relations, attributes, and logical connectives). This Set was used for extending the relational content analysis of cell biology passages from six textbooks as well as students' representations. Our findings are interesting and illustrate the insights one could gather from such an analysis. We also conducted a feasibility study to demonstrate how the method may be used in a classroom 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 contributes towards providing a coding for content analysis specific to the scientific text. The thesis highlights the importance of developing rigor by eliminating ambiguities and concludes by drawing implications of the work for research in biology education and science communication in general.

Introduction

Understanding scientific knowledge is essentially mediated by a specialized language of science and therefore understanding is interconnected with learning the language of science (Zukswert et al., 2019; Evagorou and Osborne, 2010). Research in learning science vocabulary has revealed that understanding of concepts can be enhanced by forming meaningful connections by linking words between concepts employing metacognitive strategies (Gu and Johnson, 1996; Ausubel et al., 1978; Novak, 1990; Schmitt, 2007).

It is well known in science education that students find the vocabulary of science difficult due to jargonified, isolated, scientific concepts alienating understanding of the scientific conceptions (Halliday, 2006; Lemke, 1990) and learning of scientific skills (Snow, 2010). The scientific concepts cannot be learnt in isolation, but have to be understood as a network of knowledge comprising of concepts (i.e. nodes) connected with linking words (i.e. edges) (Halliday, 2006; Minsky, 1974; Quillian, 1968).

The regular teaching and learning focus largely on developing a vocabulary of scientific concepts through their definitions presented as statements. As a result, the focus is largely 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, 2006). 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 raised some questions: Could consistent use of well-defined linking words be a good indicator of rigor in scientific language? Can rigor in scientific language be understood in terms of predicates? These questions motivated us to develop a relational content analysis 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). These questions are formulated with an underlying assumption that the meaning of any term originates from relations between other terms (network theory of meaning), and not concepts per se. 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 were paraphrased into subject-predicate-object representing a proposition. We will follow a similar syntax that is considered as [concept-linkingwords-concepts] and also as [node-link-node] in the concept mapping domain. When applied in semantic web standards this form is represented as triples of [class-objectproperty-class], [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: [concepts] for technical terms that are nouns; (relations) for verbs/linking words; (attributes) for variable properties/adjectives. Often we also underline modalities and logical connectives that define the scope of each of the statements and express reasoning forming complex sentence structures.

Identifying Research Gaps and Formulating Research Questions

While assessing the concept mapping tool and published concept maps, we identified certain inconsistencies. 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). Some of these are *meronymic inclusion, class-subclass inclusion, spatial inclusion, attributes, etc.*

Problems in Accuracy of Linking Words We began with studying the published concept maps in the public domain, published journal articles, and books. These published repositories showed findings that concept mapping 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 traditional 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. This work demonstrates its possibility. We proposed certain refinements of linking words in terms of accuracy and validating hierarchies in concept mapping.

The concept map on life in the ocean (Martin et al., 2000) as shown in Figure 2 by 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 them with the semantically well-defined linking words as shown in S2. By this, we suggest weeding out ambiguity, bringing in consistency ie 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. The traditional criteria to assign scores for the hierarchy is based on graphical and not semantic criteria. Logically, in a well-formed hierarchy, the relation must use a transitive, asymmetric, and a single linking word consistently (Mayr, 1969). In S1's 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 traditional scoring method fails to make this subtle distinction and commits a logical fallacy. This is a serious anomaly, and therefore there was a need to find an appropriate refinement in validating hierarchies in the maps. 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, 2006).

Lack of focus on process-centric propositions The verbatim use of linking words (verb based) in textbooks such as 'produces', 'supports', 'stores', 'secretes', etc. gives a colloquial sense to scientific propositions. 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 that the processes 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 place "protein traffic", into processes such as "regulation of protein traffic" and creates dynamic propositions such that the representation is to understand the "regulation of protein traffic" as a biological process vis-a-vis mere "structure of

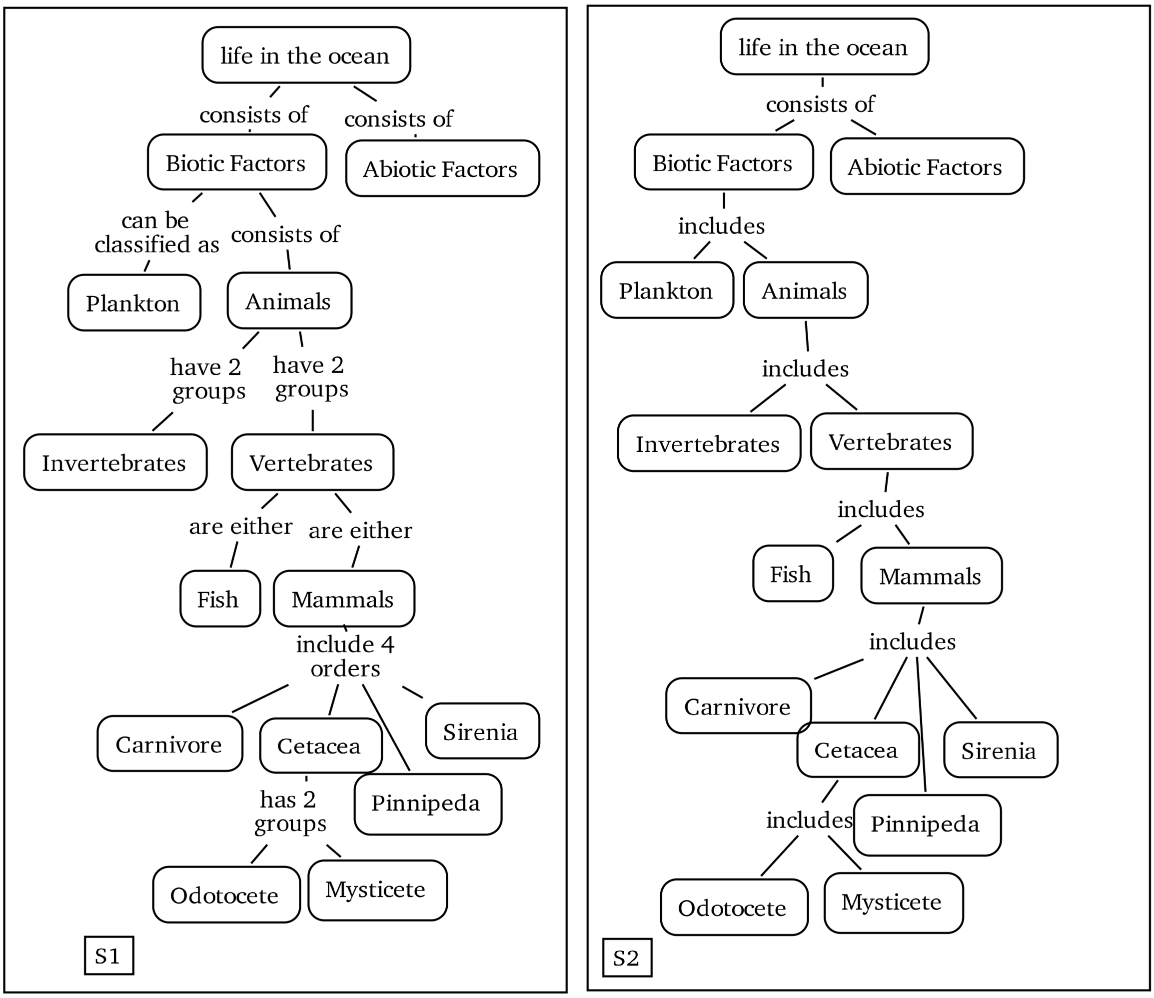

Figure 2: Concept map depicting ambiguities and inconsistencies in linking words and suggesting refinements.

protein". This enables us to focus on the question of 'how' in understanding any phenomenon.

No distinction between relations and attributes To indicate the problem of lacking 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 attributes. 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 clarification 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.

No focus on depicting change Although concept mapping tool has been used almost worldwide, it is not being used up to their full potential, ie. mostly used for depicting descriptions and not for explanations, depicting change, or thinking about dynamic systems (Canas and Novak, 2006). One of the reasons being projected was that the focus questions are more description based rather than explanation based. It is also reported that the concept maps have a strong advantage of representing static relations between concepts, but the methodology lacks the potential for representing change or dynamic relations (Derbentseva et al., 2007).

This certainly calls for re-thinking the method in order to extend the concept mapping methodology to its full potential, particularly because science is largely about giving model-based reasoning, explanations, and modeling change.

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

In scientific language, the accurate use of linking words for precision and rigor characterizes expertise (Kharatmal and Nagarjuna, 2008). Studies indicated that providing linking words helps to depict more valid propositions as compared to just providing seed concepts with the freedom to use any linking words (Kharatmal and Nagarjuna, 2009). Focusing on linking words was studied by Fisher et al. (2000) to create semantic networks in the early attempts 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 not only use appropriate linking words, but also use a diversity of linking words, while novices' usage of linking words seem often inappropriate, and their usage of similar linking words for a variety of meaning results in ambiguity (Kinchin, 2001).

As the goal of mapping knowledge is to facilitate meaningful learning, it is important 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 the gaps and hence our focus is on the re-representation of the linking words for making it explicit and rigorous, and studying if it can bear significance in achieving expertise and restructuring of knowledge. In order to address the gaps, we framed four research questions (RQ).

- RQ1: Can we design a method of representation of biology language focusing on the semantics of predicates?
- RQ2: Could this method be used for the content analysis of biology language?

- RQ3: Could this method be feasible?
- RQ4: Can predicate analysis be a measure of expertise?

To address RQ1, we begin with designing and developing a method by re-representation of linking words that lead to the modeling of biology language. To address RQ2, we use this method to demonstrate its use for the content analysis of biology textbooks and students' descriptions. RQ3 is addressed by demonstrating the feasibility of the method by comparing it with traditional methods. RQ4 is addressed by comparing the text with the expert's knowledge base to check if the re-representation of linking words can show a proximity percentage. Figure 1 in the Executive Summary provides a concise summary of the research model.

The design of the method is influenced by three major theoretical frameworks: (i) representing knowledge as a network (Minsky, 1974; Quillian, 1968), with explicit focus on semantics and drawing from the published ontologies (Ong et al., 2017; Smith et al., 2005); (ii) representational redescription model of cognitive development for a dynamic transformation between implicit and explicit knowledge (Karmiloff-Smith, 1995). We also follow various approaches for modeling biology knowledge from conceptual spaces as framework for representations based on predicate dimensions (Gardenfors, 2000), object process methodology for representing events (Dori, 2002; von Wright, 1963), systemic functional linguistics for nominalization, momenting of dynamic activities, experimentation, etc. (Halliday, 2006; Halliday and Martin, 1993; Doran and Martin, 2021; Unsworth, 2020).

As with any design and development work, we appropriated the design-based research (DBR) approach for iteratively refining the development of the re-representation method. (Brown, 1992; McKenney and Reeves, 2018). This method is evolved through features that are unique in design practices such as, creating, testing, iterative refinements, and continuous evolution of the design (Collins et al., 2004).

The Re-representation Method

The method is developed with iteration into six phases. Initially, we focused only on a preliminary set of 7 types of linking words, belonging to meronymic inclusion (part-whole), class inclusion, and other relations. We tested this by deploying it for textbook analysis. This set catered to representing structure terms in the topic. As we analyzed more textbooks from higher levels of complexity the set was necessary but not sufficient. We observed that there was a need for some more types of relations, for example, spatial inclusion, and function. This called for the iteration of the re-representation method to include other types of relations as well (iteration 1). Further, as we analyzed higher class textbooks, there was a need to introduce dynamic propositions (iteration 2) for process-based terms, and represent variable properties (iteration 3) to depict change. This further led to refining the method by adding attributes and process terms for depicting process modeling (iteration 4). As an extension, the method also integrates experimentation modeling (iteration 5). We use the iterations to represent reasoning between propositions (iteration 6).

Re-representing relations We begin by paraphrasing sentences into propositions. We focus on the linking words for their intended meaning and follow the OBO, KR research to re-represent these with well-defined predicates (Winston et al., 1987; Smith et al., 2005; Lemke, 1990; Halliday, 2006). As an example, in the two sentences, “Mitochondria have two membrane coverings” and “Mitochondria have DNA and ribosomes” the verb “have” is ambiguously used. In the first sentence, the intended meaning is about composition whereas in the second sentence, the intended meaning is about a covering. Therefore to weed out the ambiguity and provide appropriate meaning we re-represent the verbs into semantically defined predicates as “consists of” and “covered by” respectively. We followed this method which has led towards the development of a Reference Set reported in Kharatmal and Nagarjuna (2016).

Re-representing attributes We make the distinction between relations and attributes while focusing on the linking words. This is of significance 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 verbs “are of” is ambiguously used for shape as well as diameter. Since these are variable properties, these are re-represented with the predicate “has shape” and “has diameter range”. Properties enable not only rich descriptions of phenomena but also quantification in biological processes. Properties depict changes in physical qualities, or measurable quantities (Doran and Martin, 2021).

Re-representing process-centric propositions The propositions that involve physical movement, action, change of state, and causal relations are considered to be of dynamic propositions (Miller and Canas, 2008). The process terms are brought into focus by nominalization by re-representing relationships (verbs), or qualities (adjectives) into things (i.e. nouns) (Halliday, 2006). 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., 2014a). 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 has role in secretion of insulin”. Inorder to highlight the process concepts, we are using predicates such as “role of”/ “function of”, “site of” “has subprocess” etc. This is similar to “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 proposition, an important aspect in understanding biological processes.

Re-representing process as state-change We follow von Wright (1963); Dori (2002) for representing process as change in the state of objects. The states are variable properties with values of parameters. Consider the sentences, “The condensation occurs by a process of folding of the chromatin fibers. At the same time, the cell becomes spheroid, more refractile, and viscous” (DeRobertis and DeRobertis, 1985, p. 420). In prophase, the momented activities includes ‘chromatin condensation’; ‘chromatin folding ie shortening’; ‘changes in cell’. The property of each are ‘less dense’ to ‘more dense’; ‘longer’ to ‘shorter’; ‘refractile’, ‘viscous’. In the process of ‘chromatin condensation’, the object chromatin is ‘less dense’ in prior state and becomes ‘more dense’ in post state. Similarly, during ‘chromatin folding’, the object chromatin has ‘longer length’ as prior state and becomes ‘shorter length’ as post state.

Re-representing experiments As an iteration and further extension of our method, we characterize experiments following the SFL framework. Scientific explanations are characterized by nominalization, higher percentage of action verbs organized in logical sequence, whereas experiments are characterized by use of imperatives (Halliday and Martin, 1993). We analysed undergraduate students' multimodal representations of experimental design (Kharatmal et al., 2022; Martin et al., 2021) of an assay on olfactory responses of Drosophila larvae (Khurana and Siddiqi, 2013). The intermodal construal of composition and activity showing progression was observed in image and language for atleast two activities: [trapping fruit-flies (trap bottle)→(bait in bottle)→(trapped flies)] ; [setting up of single line culture (banana bottle)→(gravid female)→(standard media)→(observe)]. This was depicted as linear or cyclical with atleast 3-4 micrographs or standalone annotated image. The activity of [recording ORI] was depicted using image and verbiage with annotations. Though experiment genre may have commonalities with explanation genre, however use of property to depict rich descriptions and quantification seem to be distinctive in experiment genre with using imperatives. The property of 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. is depicted.

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

In thesis, we provide illustrations of each of these phases with examples from textbook and undergraduate students' depictions.

We now address the RQ2 and demonstrate the utility and potential of the re-representation method for content analysis method.

Using re-representation method for semantic content analysis of text

The Re-representation method is deployed to show its significance in semantic analysis of text and student's descriptions. The sentences are first paraphrased into verbatim propositions or original propositions (preserving the predicates) and then if predicates are re-written then these are considered as re-represented propositions.

Methodology, Corpus Data, Coding and Categorization

The methodology applied is an extension of the relational content analysis (Carley, 1990; Palmquist and Dale, 1997). Content analysis is comprising of “a strict and systematic set of procedures for the rigorous analysis, examination and verification of the contents of written data” (Cohen et al., 2015). The content analysis is all about making valid, replicable and objective inferences about the text on the basis of explicit rules. Content analysis is a process by which many words of texts are classified into much fewer categories (Weber, 1985). Relational analysis, although like conceptual analysis begins with the act of identifying concepts in the unit of analysis, but seeks to go beyond by exploring the relations or links between the concepts. Relational analysis is often termed as semantic analysis (Palmquist and Dale, 1997).

Corpus Data The corpus of content analysis are 6 textbooks of increasing levels of domain complexity at school, college, undergraduate, post graduate levels, on the topic of cell structure and function (NCERT, 2007b,c,a; Taylor et al., 2003; Campbell and Reece, 2008), cell division DeRobertis and DeRobertis (1985). The details of the text corpus is shown in Table 1. We adopted the textbook format for content analysis, because as a written medium it is widely used in schools and colleges. 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 topic). The corpus of textbook data are over 1100 sentences which were paraphrased into over 2100 propositions and re-represented into over 2300 propositions. In addition, we also conducted content analysis of high school students' representations and undergraduate students' representations of experimentation (Kharatmal et al., 2022).

As our interest is in the area of biology texts, we selected a very basic and fundamental topic that span from school, high school, college, to undergraduate and post graduate levels. The content for each chapter on cell biology was broken down into the number of topics, such as nucleus, mitochondria, cytoplasm, ribosomes, etc. Each topic was analyzed for sentences and the propositions as shown in Table 2.

Coding and Categorization The sentences are sampling units that are syntactical in nature. Coding units are ‘concepts’ and ‘relations’ that are contained within the sampling unit of a sentence. The concepts are technical terms (noun) in subject and object place and are domain specific whereas the predicates (verb/adjective) are not domain specific and also occur in everyday language.

Categorization of the content is crucial in the process. An exhaustive list of categories is required to address content validity. Constructing of categories involves grouping units into groups of semantic relations. For the purpose of the study, we established 3 steps of categorization wherein each coding unit was grouped into the appropriate category depending on its implicit and intended meaning.

The categorization is assigned to propositions with the form subject-predicate-object. The first step is to categorize the predicate based on the decision whether it belongs to relation or attribute (step 1). The relations connect the subject and

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 (DeRobertis and DeRobertis, 1985)	Chapter 15: Mitosis and Cell Division	420-423

Table 1: Details of Text Corpus.

the object place that are both concepts. Whereas the attributes connect subject with object that describes property of the subject (variable property) for example as color, shape, size, form, temperature, intensity, etc.

When the predicate belong to relations, these are further categorized as step 2 based on the role and dimension as: meronymy inclusion, class inclusion, spatial inclusion, functional, process. While the attributes are also categorized as step 2 in terms of: form, measurable property, physical quantity, so on so forth.

Further for step 3, the predicates are categorized into specific categories for relations as well as attributes. For example, meronymy relations includes part-whole and material-composition; the spatial inclusion relation includes location and region. Although there are 3 steps of categorization, however during content analysis, all these are assigned simultaneously. The Figures 3 and 4 shows the 3 steps of categorization followed for relations and attributes respectively for all the corpus. We assign the categories to the verbatim propositions and re-represented propositions.

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 2: Total corpus data from textbooks.

Data Analysis, Findings and Discussion

The main essence of content analysis using the developed method, has been for seeking patterns, categories, themes and regularities from school, college, and undergraduate textbook data (Cohen et al., 2015).

After the 3 steps of categorization, each of the predicates are collated dimension-wise, counted for the kinds and their frequency. This is followed for each topic and then a cumulative analysis is done for each step, on both verbatim propositions and re-represented propositions. This is done to compare the before and after re-representation changes.

We first briefly illustrate the outcome and patterns from the content analysis.

Outcome and Patterns

Using the re-representation method, the entire content analysis emerged into an outcome indicating categories that are sufficient to represent the cell biology domain. These are: (i) Six kinds of categories for relations – meronymic, class-subclass, spatial, function, process, causal. (ii) Three kinds of categories for attributes – form, measurable, physical quantities. (iii) Four ontological categories are sufficient to represent the cell biology domain – structure based, class based, process based, attribute based. (iv) A Reference Set of predicate vocabulary of relations and attributes is shown in Tables 3 and 4 respectively.

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

Table 3: Dimension-wise categorization of predicates (relations).

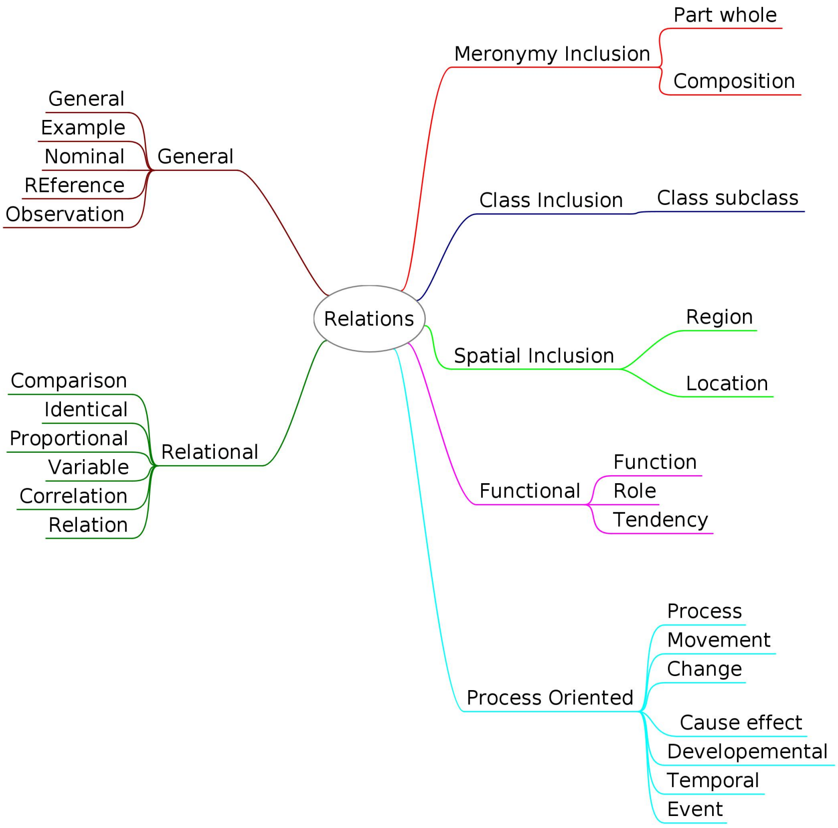

Figure 3: Categorization of Relations upto 3 Steps.

As a result of entire content analysis of verbatim propositions and re-represented propositions, the following patterns have emerged. These are: (i) As the concepts increases with increasing complexity of texts, but the relations attain saturation or constancy. This indicates that even if new knowledge is introduced to the domain it may add new concepts but the same relations would be (re)used (ii) Increase in attributes with the increasing complexity of the text, indicating that higher level or undergraduate level texts are richer in descriptions of variable properties, giving scope for quantification, parameterization and modelling. (iii) When comparing the school and college level texts, there is an increase in process based relations in college level texts. (iv) Among the six kinds of relations categories, the function category showed prominence, perhaps due to the structure-function relation of the domain itself. (v) Due to nominalization, the function category of relations attained saturation. (vi) Among the three kinds of attributes categories, the form category showed prominence. Focusing on predicates also demonstrates restructuring or conceptual change (Thagard, 1992; Nersessian, 1998). We could characterize the re-representation of predicates into – category shift, coalescence, differentiation (Carey, 1987).

As our focus has been on the predicates, we demonstrate its significance in science education research by way of two studies – feasibility (RQ3) and proximity (RQ4).

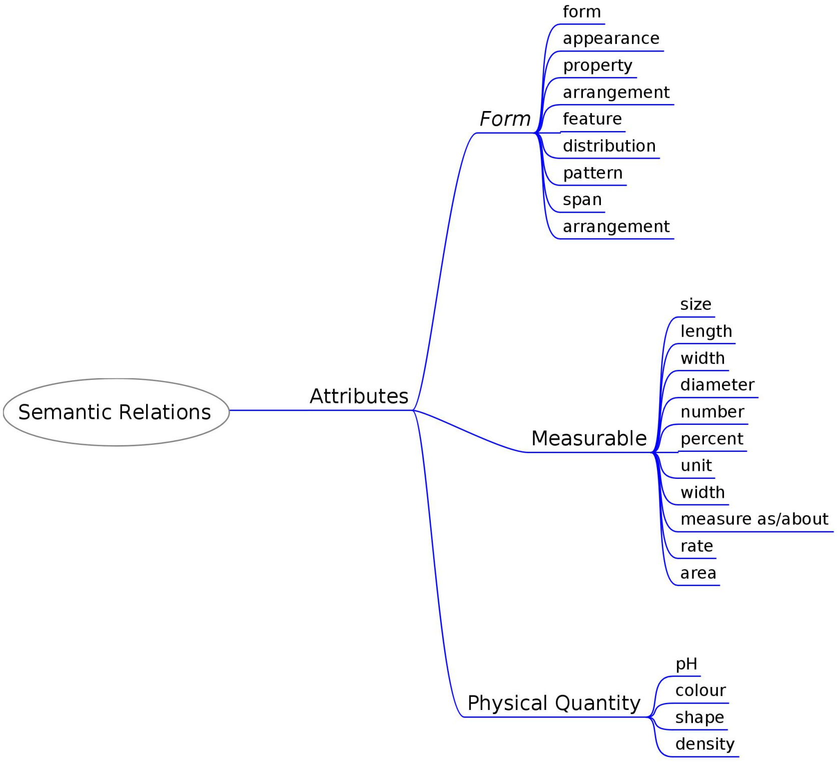

Figure 4: Categorization of Attributes upto 3 Steps.

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 4: Dimension-wise categorization of predicates (attributes).

Feasibility of using a minimal set of relations

We conducted a study to find out, if a minimal set of linking words are provided, would it be difficult or easy for students to come up with a consistent representation, without any loss of expression. The study had three groups of a homogeneous sample of students (age 13-14 years, mixed gender) studying in grade IX from a local urban school. The three groups (in three different classrooms) were assigned with three different modes of representations – description, traditional concept mapping, re-represented concept mapping. The chapter on “The Fundamental Unit of Life” from grade IX 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”. Figure 5 shows a re-represented concept map on nucleus with a minimal set of linking words.

While the description method and traditional concept mapping used 30 and 42 kinds of linking words respectively during accomplishment of the task, the same domain was represented with only 7 linking words provided using the re-represented concept mapping. For all the three groups, a score of 1 each was assigned to each non-redundant concept and a score of 1 each for valid relation of nucleus and mitochondria. The data were statiscally analyzed. A single factor analysis of variance (ANOVA) was used to compare the variance of the three groups. For the concepts of nucleus the F(2, 89, 91) = 2.66, $p>0.05$ was found non-significant. However for linking words of nucleus the F(2, 89, 91) = 8.20, $p<0.05$; and for the concepts of mitochondria the F(2, 89, 91) = 4.13, $p<0.05$; and for the linking words of mitochondria the F(2, 89, 91) = 4.50, $p<0.05$ were found to be significant.

In order 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 re-represented 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 concepts (nodes), there has been no significant differences in the traditional concept mapping method t(56.52) = 1.75, $p>0.05$ and the re-represented concept mapping method t(57.02) = 0.8, $p>0.05$, which shows that the re-represented concept mapping does not affect the representation of critical concepts.

The study demonstrated that even a minimal set of linking words (constrained concept mapping) does not hinder the representation of critical concepts, rather it serves as a scaffolding, without any loss of expression (or knowledge) in the re-representation method. This study is published in Kharatmal and Nagarjuna (2009).

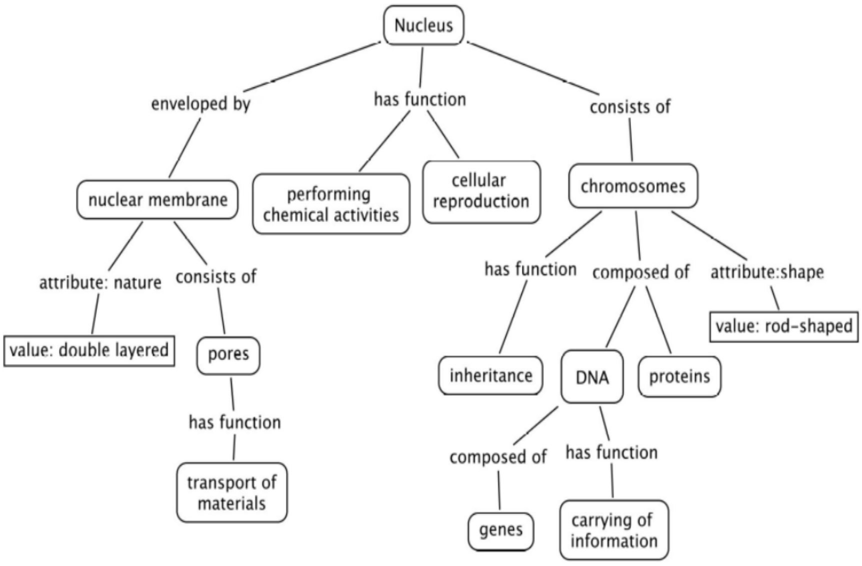

Figure 5: A re-represented concept map on ‘nucleus’ created using the minimal set of linking words.

Proximity percentage as an indicator of expertise

We address the proximity analysis by studying if the choice of predicates characterizes expertise. The topics on ‘nucleus’ and ‘mitochondria’ were analyzed from three books UG-1 (Taylor et al., 2003), UG-2 (Campbell and Reece, 2008), PG-1 (DeRobertis and DeRobertis, 1985). The proximity index is calculated as the number of common predicates between the verbatim and re-represented predicates divided by the number of verbatim predicates for a topic. The proximity (or overlap) of the verbatim predicates with the Reference Set increases with UG-1 to UG-2 to PG-1. The UG-1 level text already showed higher proximity of 83 %, though not so dramatic. However, this trend is more apparent when the same analysis is carried out for school texts where the proximity was found to be only about 50 %. This study demonstrates that proximity of predicates used in a text with those of the Reference Set correlates with expertise. This implies that with applying the re-represented relations, the text becomes closer to an experts representations and bears an indicator of expertise. This study is published in Kharatmal and Nagarjuna (2016).

We also compared the school text with expert’s text for the choice and usage of predicates. A 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 level texts usage of verbatim choice of predicates as well as re-represented predicates with 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.

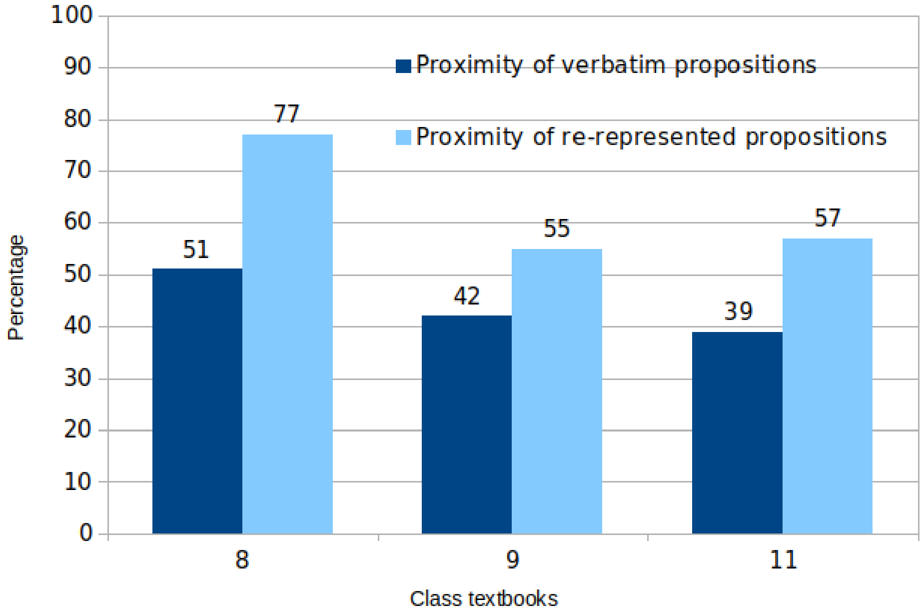

Figure 6: Greater proximity percentage of re-represented predicates with expert’s text.

We observed the proximity percentage increased by a range of atleast 10-20 %. There is a higher degree of proximity of re-represented predicates as compared from the verbatim predicates as shown in Figure 6. This showed that certainly if and when we do re-representation it gets much closer to an experts representations. This indicates that characterizing representations using predicates can become an indicator of expertise. In similar way, students' representations can be compared with expert's representations to study novice-expert transformation.

In another method, we compared the re-represented predicates with the expert's repository of the published relations and attributes (Ong et al., 2017). One important aspect in doing this comparison, is also to validate the re-representation method. Each of the verbatim and re-represented predicates for 5 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 7.

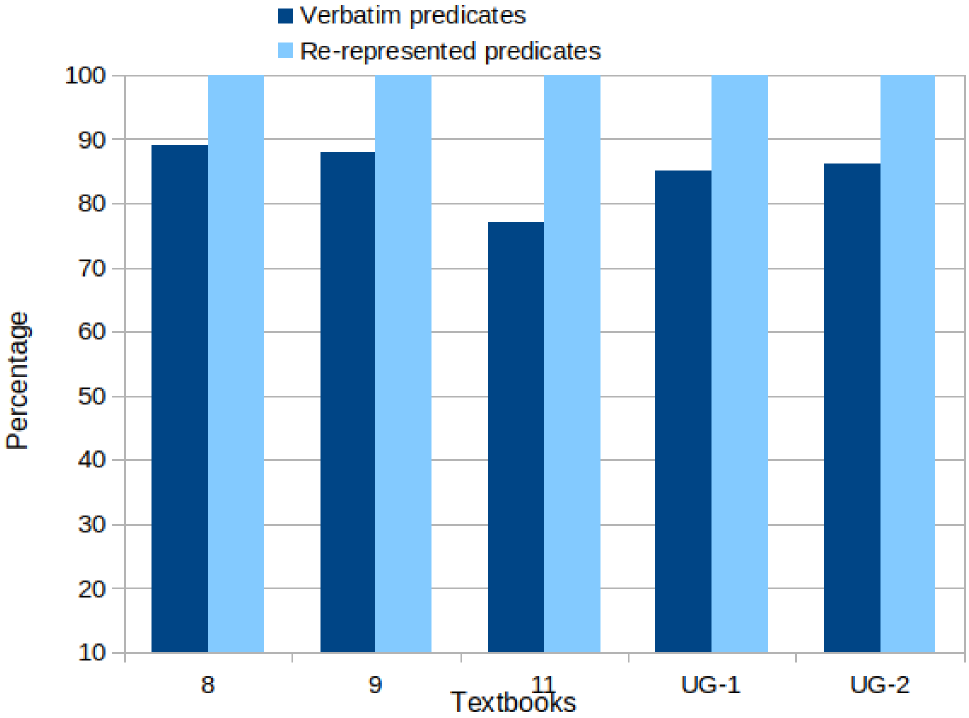

Figure 7: Higher proximity percentage with expert's published repository.

Illustrations of the Re-represented Method

We illustrate the relational content analysis with a sample 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 below. 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. A snippet of three examples is provided from the entire illustrations provided in the thesis.

1. Other types of membrane lipids <u>are</u> also amphipathic.
2. The leucoplasts <u>are</u> the <u>colorless</u> plastids <u>of</u> varied <u>shapes</u> and <u>sizes</u> <u>with</u> stored nutrients.

3. 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:

(1′) Other types of membrane lipids also have amphipathic property [R2, C9].

(2′) The leucoplasts are [R1, C2] the colorless plastids. They have various shapes and sizes [R3, C7, C8] and contain [R4, C4] stored nutrients.

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

In the sentences from (1) to (3), the verbs is/are are used in four different senses: class inclusion, attribute, 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 (Kharatmal and Nagarjuna, 2010). Therefore, it calls for re-representation.

In sentence (1), the verb are is being used for the intended meaning of attribute. For all attributes, 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 (2) 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 attribute. While in (iii) the preposition with is used to express the containment relation. 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 (3), the intended meaning is about describing the shape, diameter, and length as attributes. 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.

Compiled List of Rules and Categories

The list is not exhaustive even for a given domain. This serves the limited purpose of demonstrating the potential 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:

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 [role/has function]

C6 Process [changes into, develops from, becomes, develops, moves towards, before, results into]

Attributes:

C7 Geometric Form [shape, appearance]

C8 Measurable [size, length, diameter, width]

C9 Quality [color]

Re-representation *rules*:

R1 Pass those statements that already use the relations and attributes 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 relation and attributes 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 relation or attributes as appropriate.

R4 Eliminate the lone usage of prepositions as semantic relations between subject and object, and use them only in combination with the relation and attributes available in the Reference Set and to create complex expressions between class terms.

R5 Rewrite the adjectives in the form of attributes 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 relations such as has role/has function, has sub-process, is site of, etc.

R7 Rewrite the verbs in the form of relations 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.

Conclusions, Limitations and Implications

One of the main contributions of our work is providing coding for content analysis specific to scientific text. This can be used by researchers in language of science. We hope that it might enhance the consistent use of vocabulary even in subject like biology. A controlled natural language vocabulary is possible and it can enhance rigor in science communication.

Our work involved a fully manual (supervised) method for creating a predicate vocabulary from biology text arriving at 6 categories of relations and 3 categories of attributes. Therefore, generating a vocabulary of predicates for each domain is possible, and it can be realized from the published repository of predicates.

In knowledge organization, Holley and Dansereau (1984) in their study on using networking as a spatial learning strategy, identified six relations – part of, type of, leads to, similar to, has characteristics, indicates/illustrates. Faletti and Fisher (1996) using SemNet in biology, reported 3 relations – set/member, whole/part, and characteristic – being used more frequently than other relations. In one of the studies on concept mapping, linking words such as – is measure of, has property of, depends on, is a form of, is mass divided by, divided by volume, equals were provided, and it was established that the scoring of maps in this case was straight- forward, and hence could be applied for large scale assessment (Yin et al., 2004). A form of concept map, called a knowledge integration map has been deployed in an online inquiry-based learning unit where it is suggested that students in a peer-review condition focused more on links and linking words (Schwendimann and Linn, 2016).

Various studies have provided evidence emphasizing linking words (linking phrases) in undergraduate biology. A research study compiled a list of linking words based on their reflections of teaching and extracting from books to include in undergraduate introductory biology courses (Javonillo and Martin-Dunlop, 2019). These linking words describe quantitative, structural, and temporal statements taught in introductory biology. 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, 2020). The quantifying of the linking words resulted in a higher number of dynamic linking words, indicating the prominence of processes. This is also similar with our findings.

In the semantic spectrum, the number of predicate terms used in formal expressions is fewer than those in lesser formal expressions (McGuinness, 2003). In Controlled Natural Language (CNL), researchers have used a constrained set of well defined predicates for creating unambiguous relations (Clark et al., 2010). Knowingly or unknowingly the value of minimalism of predicate terms is promoted in rigorous subjects. Given the possibility of predicate vocabulary, we can rewrite (rigorous) scientific text.

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 a higher level biology textbooks (Chaudhri et al., 2014b; Koć-Januchta et al., 2020). The method of relational content analysis can be applied for comparison of textbooks as well. 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.

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, may not occur. We suggest that when text is written as nuclear predicates, then science text becomes accessible to ESL learners.

Creating a controlled vocabulary of biology can reduce the burden of learning the heavy jargon for biology students (Bennett et al., 2013). Towards this, 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 predicates as metadata for explicit statements applying the method of Aristotelian definition. In their illustration of a glossary of one such term about eukaryotic cell, the relations being used are: has part as relation and size, diameter as attributes. Our work of relational content analysis along with the compiled list of Reference Set can facilitate as a precursor to creating such a knowledge base.

Our work can be considered parallel with the efforts of at least two forms of CNL in Attempto Controlled English (ACE) (Fuchs et al., 2008; Schwitter, 2010; Kuhn, 2010, 2014), Computer Processable Language (CPL) (Clark et al., 2010). Although ACE is applied to the English language, our proposed method of re-representation 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 predicates extracted from the content analysis method can bear application for the automated reasoning algorithm. The set of predicates can also be validated by the recent open AI applications for example chatGPT that uses large language model as an algorithm to process natural language.

Lastly, the future work can be in the direction of using the method for creating intelligent textbooks, 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 and designing games that facilitate the use of writing text with 'nuclear predicates'. It may even become a criterion for the assessment of the written text.

The work is limited to content analysis of textual descriptions and of students' writings and drawings. The work so far does not address the activities, images, graphs, equations, figures, box items, questions, historical snippets, etc.

Organization of the Thesis

Chapter 1: Introduction, Motivation and Objective The motivations for the thesis drives from semantic network (meaning is in the network) area, language of science, educational psychology, implicit-explicit knowledge, semantic principles for linking words.

We propose that the re-representation method developed with the reference set of well defined linking words can act as interface for facilitating implicit knowledge to explicit knowledge.

Chapter 2: Literature Review and Theoretical Framework The thesis applies theoretical frameworks from multiple dimensions for designing and developing the re-representation method. The chapter presents a discussion about the insights borrowed from three theoretical framework. The re-representational redescription model for implicit to explicit, semantic network for semantic holism, structure function linguistics for nominalization, conceptual spaces for predicates, etc. The chapter also presents a literature review of the various theoretical frameworks, reviewing other research works that is focusing on the linking words, and also identifying gaps. This lead 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. Representing process-centric propositions by nominalization (Halliday, 2006)
3. Making distinction between relations and attributes ie variable properties (Ong et al., 2017; Smith et al., 2005; Gardenfors, 2000)
4. Modelling processes (scientific explanations) as state-change using variable properties (von Wright, 1963; Sowa, 2003; Dori, 2002; Doran and Martin, 2021)
5. Modelling experiments as momented activities and variable properties (Doran and Martin, 2021; Unsworth, 2020; Halliday and Martin, 1993)
6. Reasoning between propositions for reasoning between statements using logical connectives (Copi et al., 1972)

   The chapter also provides an illustration of re-representation rules and coding used in content analysis of scientific text.

Chapter 4: Re-representation method for semantic analysis of text In this chapter, the re-representation method is deployed to show its significance in semantic analysis of text. In this chapter, we present the details of textbook analysis and students' analysis following the relational content analysis method. We present the coding and categorization of data with illustrations. The chapter ends with framing of research questions and making connections with the empirical studies for each of them.

Chapter 5: Data Analysis, Findings and Discussion This chapter presents a detailed analysis leading to the patterns in the growth of knowledge, using proximity as indicator for expertise. We also discuss high school students' findings of feasibility study. Further, we also characterize experimentation using undergraduate biology students' intermodal data analysis of experimentation. We also illustrate the rules and coding suggested for content analysis specific to scientific text with examples.

Chapter 6: Conclusion and Implications The conclusions discussed in the context of the work and the implications are discussed in the area of concept mapping, textbook analysis, biology education research, etc.

Peer-reviewed publications arising from the work in this thesis

1. 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

2. 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. Springer. DOI: 10.1007/978-3-319-45501-3_25
3. Meena Kharatmal & Nagarjuna G. (2010): Introducing rigor in concept maps. In M. Croitoru, S. Ferre, and D. Lukose (Eds.), Lecture Notes in Artificial Intelligence: Vol. 6208. International Conference on Conceptual Structures 2010: From Information to Intelligence (p. 199-202). Berlin, Germany: Springer-Verlag. Doi: 10.1007/978-3-642-14197-3_22
4. Meena Kharatmal (2009): Concept Mapping for Eliciting Students' Understanding of Science. Indian Educational Review, 45(2), pp.31-43.

Publications from International Conference Presentations and Proceedings

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.

Bibliography

D. Ausubel, J. Novak, and H. Hanesian. *Educational psychology: a cognitive view*. Holt, Rinehart and Winston, 1978. ISBN 9780030899515.

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.

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, April 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):e66–73, July 2006. ISSN 1367-4811. doi: 10.1093/bioinformatics/btl266.

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

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

Susan Carey. *Conceptual Change in Childhood*. Mit Press, 1987. ISBN 9780262530736.

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.

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, 2014a. doi: 10.3233/978-1-61499-438-1-263.

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, 2014b. doi: 10.3233/978-1-61499-438-1-415.

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_1$0.

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.

L. Cohen, L. Manion, and K. Morrison. *Research Methods in Education.* Routledge, New York, seventh edition, 2015.

Allan Collins, Diana Joseph, and Katerine Bielaczyc. Design Research: Theoretical and Methodological Issues. *Journal of the Learning Sciences*, 13(1):15–42, January 2004. ISSN 1050-8406. doi: 10.1207/s15327809jls1301_2.

Irving M. Copi, Carl Cohen, and KD (Kenneth D.) McMahon. *Introduction to logic.* Macmillan New York, 1972.

J. V. Costa, F. E. Rocha, and E. 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.

Natalia Derbentseva, Frank Safayeni, and J. Cañas, Alberto. 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`.

E. D. P. De DeRobertis and E. D. P. De DeRobertis. *Cell and Molecular Biology.* Lea & Febiger, New Delhi, India, 1985.

Yaegan Doran and J Martin. Field relations: Understanding scientific explanations. In K. Maton, J. R. Martin, and Y. J. Doran, editors, *Teaching Science: Language, Knowledge, Pedagogy.* Routledge, 2021.

D. Dori. *Object-Process Methodology.* Springer-Verlag, New York, 1 edition, 2002. ISBN 9783540654711.

E. Erdem and R. Yeniterzi. Transforming Controlled Natural Language Biomedical Queries into Answer Set Programs. In *BioNLP@HLT-NAACL*, 2009. doi: 10.3115/1572364.1572381.

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.

Joseph Faletti and Kathleen M. Fisher. The information in relations in biology, or the unexamined relation is not worth having. In Kathleen M. Fisher and Michael R. Kibby, editors, *Knowledge Acquisition, Organization, and Use in Biology*, NATO ASI Series, pages 182–205. Springer, 1996. ISBN 978-3-642-61047-9. doi: 10.1007/978-3-642-61047-9$_1$3.

K. M Fisher, J. H Wandersee, and D. E Moody. *Mapping biology knowledge*, volume 11. Springer, 2000.

Marta Fonseca, Beatriz Oliveira, Pedro Carreiro-Martins, Nuno Neuparth, and António Rendas. Revisiting the role of concept mapping in teaching and learning pathophysiology for medical students. 44(3):475–481, 2020. ISSN 1043-4046. doi: 10.1152/advan.00020.2020. URL `https://journals.physiology.org/doi/full/10.1152/advan.00020.2020`. Publisher: American Physiological Society.

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`.

Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. Attempto controlled english for knowledge representation. In *Reasoning web*, pages 104–124. Springer, 2008.

Peter Gardenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, USA, 2000. ISBN 978-0-262-57219-4.

Y. Gu and R.K. Johnson. Vocabulary learning strategies and language learning. In *Language and Learning*. Wiley International., 1996.

M. A. K. Halliday. *The Language of Science*. Continuum International Publishing Group, London, 2006. ISBN 9780826488275.

M. A. K. Halliday and J. R. Martin. *Writing Science: Literacy and Discursive Power*. The Falmer Press, London, 1993. ISBN 9780750700467.

M. A. K. Halliday and C. M. I. Matthiessen. *An Introduction to Functional Grammar*. Routledge, London, 2004.

CHARLES D. Holley and DONALD F. Dansereau. Networking: The technique and the empirical evidence. In CHARLES D. Holley and DONALD F. Dansereau, editors, *Spatial Learning Strategies*, pages 81–108. Academic Press, 1984. ISBN 978-0-12-352620-5. doi: 10.1016/B978-0-12-352620-5.50010-2. URL `https://www.sciencedirect.com/science/article/pii/B9780123526205500102`.

Robert Javonillo and Catherine Martin-Dunlop. Linking phrases for concept mapping in introductory college biology. 45(3), 2019. URL `https://www.academia.edu/65349593/Linking`$_{P}hrases_{f}or_{C}oncept_{M}apping_{i}n_{I}ntroductory_{C}ollege_{B}i$

Ingvar Johansson. Four Kinds of Is_a Relation. In *Applied Ontology: An Introduction*, pages 235–254. De Gruyter, May 2013. ISBN 9783110324860.

Annette Karmiloff-Smith. *Beyond Modularity: A Developmental Perspective on Cognitive Science*. The MIT Press, September 1995. ISBN 9780262611145.

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, 2006.

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.

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.

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.

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.

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.

Meena Kharatmal, Mayur Gaikwad, Aashutosh Mule, and Jaikishan Advani. Exploring college biology students' understanding of experimental design. 14:2–7, 2022. doi: 10.57443/ajbe.14.0_2.

Sukant Khurana and Obaid Siddiqi. Olfactory responses of drosophila larvae. 38 (4):315–323, 2013. ISSN 0379-864X. doi: 10.1093/chemse/bjs144. URL `https://doi.org/10.1093/chemse/bjs144`.

I. M. Kinchin. Can a novice be viewed as an expert upside-down?. 83(303):91–95, 2001. URL `http://eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ641852`.

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.

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.

J. L. Lemke. *Talking Science: Language, Learning, and Values.* Praeger, Norwood, N.J, January 1990. ISBN 9780893915667.

B. L. Martin, J. J. Mintzes, and I. 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`.

J. R. Martin, L. Unsworth, and D. Rose. Condensing meaning: Imagic aggregations in secondary school science. 2021.

E. Mayr. *Principles of systematic zoology.* McGraw-Hill New York, 1969. URL `http://hymenoptera.tamu.edu/courses/ento601/pdf/mayr`$_a shlock_c hapter_2.pdf$.

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.

N. Miller and P. Canas. Effect of the nature of the focus question on presence of dynamic propositions in a concept map. 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*, Tallinn, Estonia & Helsinki, Finland, 2008.

Marvin Minsky. A Framework for Representing Knowledge. Technical Report Memo 306, MIT-AI Laboratory, USA, 1974.

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.

N. J Nersessian. Conceptual change. In W. Bechtel and G. Graham, editors, *A companion to cognitive science*, pages 155–166. Blackwell, 1998.

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.

OBOFoundry. The Open Biological and Biomedical Ontologies, 2020. URL `http://www.obofoundry.org/docs/OperationsCommittee.html`.

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.

K.M. Palmquist, M. E. Carley and T.A. Dale. Two applications of automated text analysis: Analyzing literary and non-literary texts. In C. Roberts, editor, *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Tanscripts.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1997.

M. R. Quillian. Semantic Memories. In M. 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.

Norbert Schmitt. Current perspectives on vocabulary teaching and learning. *International handbook of English language teaching*, pages 827–841, 2007.

Beat A. Schwendimann and Marcia C. Linn. Comparing two forms of concept map critique activities to facilitate knowledge integration processes in evolution education. 53(1):70–94, 2016. ISSN 1098-2736. doi: 10.1002/tea.21244. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/tea.21244`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/tea.21244.

Rolf Schwitter. Controlled natural languages for knowledge representation. In *Coling 2010: Posters*, pages 1113–1121, 2010.

B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L Rector, and C. Rosse. Relations in biomedical ontologies. *Genome biology*, 6(5):R46, 2005.

Catherine E. Snow. Academic language and the challenge of reading for learning about science. 328(5977):450–452, 2010. doi: 10.1126/science.1182597. URL `https://www.science.org/doi/10.1126/science.1182597`. Publisher: American Association for the Advancement of Science.

J. F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations.* Brooks / Cole, USA, 001 edition, 2003. ISBN 0534949657.

John Sowa. Concept mapping. San Fancisco, USA, 2006. URL `http://www.jfsowa.com/talks/cmapping.pdf`.

D. J. Taylor, N. P. O. Green, and G. W. Stout. *Biological Science.* Cambridge University Press, UK, 3 edition, 2003. ISBN 0521561787.

Paul Thagard. *Conceptual Revolutions.* Princeton University Press, 1992. ISBN 9780691024905.

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 Sys-*

temic Functional Linguistics, pages 131–151. Bloomsbury Publishing., 2020. URL `https://doi.org/10.5040/9781350116092.ch-007`.

Georg Henrik von Wright. *Norm and action: a logical enquiry.* Humanities, 1963.

R. P. Weber. *Basic content analysis.* Sage, New Delhi, 1985.

M. E Winston, R. Chaffin, and D. Herrmann. A taxonomy of part-whole relations. *Cognitive science*, 11(4):417–444, 1987.

Y. Yin, J. Vanides, M. A Ruiz-Primo, C. C Ayala, and R. J. Shavelson. A comparison of two construct-a-concept-map science assessments: Created linking phrases and selected linking phrases. pages 1–28, 2004.

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.