APPLYING COMPUTATIONAL METAPHOR IDENTIFICATION TO MIDDLE SCHOOL STUDENTS' WRITING ABOUT CELLULAR REPRODUCTION

Metaphors allow students to grasp abstract concepts they cannot touch or see directly, but they can also lead to cognitive constraints and misconceptions. Fostering an awareness of such metaphors is a crucial step in developing expertlike, flexible scientific conceptual understanding. This paper describes a novel text analysis system for identifying potential metaphors implicit in students' writing. Computational metaphor identification (CMI) was integrated into a webbased inquiry science module to analyze seventh grade science students' responses to a series of questions about the processes of cellular reproduction (mitosis and meiosis). Based on the hypothesis that people regularly link bodies and buildings metaphorically (Lakoff, Espenson, & Schwartz, 1991), CMI was used to look for metaphors framing concepts of cellular reproduction in terms of concepts from the domain of architecture. The computationally identified metaphors indicate that students may be using a variety of metaphors, for example, a cell is like a building. These results suggest that CMI could be a powerful tool, not only allowing teachers and researchers to determine what metaphors students are using, but also encouraging students to reflect critically on the metaphors they use and consider potential alternatives.

Eric Baumer, Department of Informatics, University of California, Irvine Lindsey Richland, Department of Education, University of California, Irvine Bill Tomlinson, University of California, Irvine

Metaphor in Scientific Thinking

Lakoff and Johnson (1980) analyze English language patterns to argue that metaphor is not just a poetic or literary device, but rather is a fundamental aspect of how humans conceptualize and learn about the world. For example, *money*, an abstract concept, is often referred to as if it were a *liquid*: "they froze his assets," "I poured all my money into the stock market," "her cash flow increased," "his funding dried up." The deeply embedded nature of this metaphor is evident in the prevalence of these language structures. Metaphor is similarly crucial to scientific thinking, both in science practice (Blanchette & Dunbar, 2001; Gordon, 1974) , and in science learning, e.g., (D. Gentner & D.R. Gentner, 1983; Zook & Di Vesta, 1991) . In classic analyses, Gordon (1974) discusses how a number of scientific and technological discoveries and inventions, including the printing press, the laws of planetary motion, and the telephone, were all greatly influenced by metaphorical thinking. Gentner and Gentner (1983) experimentally identify two common metaphors for electrical circuits: water flowing through a pipe, and a crowd running around a race-track. Importantly, they reveal how each of these two metaphors, in certain situations, can lead to different conclusions about the current and voltage in different circuit configurations. Thus in some problem contexts one metaphor is useful while the other may constrain students' thinking and lead to potential misconceptions. Such misconceptions do not mean that the use of metaphor is to be avoided. Rather, improving students' facility with noticing, critically examining, and creatively generating different metaphors thus can help students avoid potential misconceptions while leveraging the power of metaphorical thinking in science.

To help foster students and teacher's ability to notice and flexibly use such metaphors, the authors are developing a system for computational metaphor identification (CMI), which analyzes large bodies of text to identify linguist patterns indicative of potential metaphors. This paper presents a summary of the CMI technique and an application thereof to the analysis of seventh grade students' writing in response to questions about mitosis and meiosis. The results presented here indicate that CMI has significant potential as a powerful educational tool, both for students and for teachers.

Computational Metaphor Identification

This section provides a high-level summary of the computational techniques involved in CMI, which extend previous work in computational linguistics (REF Mason).

Metaphors are conceptual mappings wherein a concept from a source domain partially structures the understanding of a concept from a target domain. In the example cited above about liquids and money, the target concept *money* is partially framed in terms of the source concept *liquid*. CMI begins by gathering corpora for the source and target domains. In this paper, the target domain is that of students' thinking about cellular reproduction. Thus, samples of students' writing about mitosis and meiosis were collected (see more about data below). For the source corpus, Wikipedia articles are used, as they provide a readily available, large source of content on a wide variety of topics. A source corpus for a given domain consists of all the Wikipedia articles in that category as well as all articles in its subcategory. All documents in the source and target corpora are automatically parsed (Klein & Manning, 2003; de Marneffe, MacCartney, & Manning, 2006) to extract grammatical relationships between nouns and verbs, such as determining a verb's subject, direct object, or indirect object.

The crux of CMI is selectional preference learning (Resnik, 1993) ,which identifies the tendency of particular words to appear with certain other classes of words in specific grammatical relationships. For example, words such as "water," "oil," or "coffee" tend to be the direct object of "pour," or the subject of "flow," "freeze," or "evaporate." Using the parsed documents, CMI calculates selectional preferences of the characteristic nouns in a corpus, where characteristic means that the noun is highly frequent in the corpus relative to its frequency in general English (Kilgarriff, 2003) . Thus, we can quantify the degree to which, e.g., "water" selects to be the direct object of "pour." Individual nouns are aggregated into classes of words that they might represent using WordNet (Fellbaum, 1998) , an ontological dictionary. WordNet uses synsets, which are sets of synonyms, to describe classes of words. For example, the words "beverage," "drink," "drinkable," and "potable" make up a synset for the concept of any liquid that is suitable for drinking. These word synsets are then clustered based on the verbs for which they select, where synsets that select for similar verbs are grouped together to form a cluster of conceptually related synsets.

To identify metaphors, CMI looks for mappings from clusters in a source corpus to those in a target corpus. For example, in a corpus about Water, a cluster for *liquid* would select

for the verbs "pour," "flow," "freeze," and "evaporate" in various grammatical relationships. In a corpus about Finance, the cluster for *money* would also select for many of those same verbs, e.g., "she *poured money* into his savings," or "they *froze* his *assets*." Based on the systematicity of these correspondences, each mapping is given a confidence score to indicate how likely the linguistic patterns are to evidence a conceptual metaphor. One of the strengths of CMI is that it works in the aggregate. While individual instances of phrases like "freezing assets" and "pouring funds" may not at first glance appear metaphorical, it is the systematicity of these patterns that becomes compelling evidence for the existence of a metaphor.

An important aspect of CMI is that it identifies only linguistic patterns potentially indicative of conceptual metaphors, not the metaphors themselves. As Lakoff (Lakoff, 1993) emphasizes, metaphor is primarily a cognitive phenomenon wherein understanding of one concept is structured partially in terms of another, and that metaphorical uses of language are instantiations of conceptual metaphors that serve as evidence for the cognitive phenomenon. CMI leverages computational power to search through large bodies of content, identifies patterns of potential interest, and presents those patterns to a human user for interpretation.

Identifying Metaphors

To test its efficacy in identifying metaphors, CMI has been applied to analyze seventh grade students' writing about cellular reproduction. Informed by prior work describing the metaphor that BODIES ARE BUILDINGS (Lakoff et al., 1991) ,the text was analyzed for the presence of metaphors that mapped from the domain of Architecture, referring to the design and construction of buildings.

Data

The data analyzed for metaphor usage were collected as part of a study on children's learning from a Web-based Inquiry Science Environment (WISE, http://wise.berkeley.edu) module teaching about cellular reproduction concepts. Students answered some prompts before any instruction, and then throughout instruction they were given interactive questions. Questions before and during instruction were open-ended, such as: "What are some differences between mitosis and meiosis? List as many as you can" or, "Do you think offspring of ALL organisms are always different from their parents? Why or why not?" Their response data as a whole were analyzed for metaphors that were underlying the way they understood and wrote about the mitosis and meiosis concepts. This corpus included 71 students' answers to 21 different questions, totaling 2,054 sentences with 19,865 words. For the source corpus, Wikipedia articles were gathered from the Architecture category and its subcategories. This corpus includes 3,478 articles, totaling 74,803 sentences with 1,544,870 words.

Results

Conf	Source	Target Examples	Source Examples
5.16	material	" created from an already existing cell "	"creating electricity from hydrogen"
		"mixing dna with another organism"	"mixed with an aggregate"
		"producing more identical cells"	"produces synthetic Gypsum"
5.02	style	"cell is created"	"style of shopping center was created"
		" combine with the original daughter cell "	"combine it with other styles"
		"leads to new organisms"	"led to the Baroque style "
4.84	building	"creating one cell"	"created blob-like architecture"
		"produces completely alike organisms"	"produced Beaux-Arts architecture"
		" created from an already existing cell "	" creates isovists from building "
4.48	piece	"replicated in each new cell"	"replicated in red brick tile"
		"cell is created"	"tiles were created"
		" made up of mostly multiple cells "	"made of two pieces"
4.15	structure	"creating one cell"	"create decorative ornaments"
		"producing more identical cells"	"produced architecture"
		"divides into two cells"	"divided into two arches"
3.58	design	"organisms are created"	"designs were created"
	0	"produce four daughter cells"	"producing a design"
		"made from two mixed cells"	"made from the design"
3.54	water	"mix with other different cells"	"mixed with water"
		"combine with the original	"combining with water"
For rea	asons of sp	bace, this results section focuses on the " made up of one cell "	the computationally identified " made of red lime "

metaphors for the concept of *cell*, that is, potential metaphors that students might use to frame their understanding of cells. Other potential analyses, for example around the concepts of *DNA* or *meiosis*, are beyond the space limits here. Table 1 presents the upper one percentile of potential metaphors for cell based on confidence score, i.e., the likelihood that the text identified is pointing to a metaphor rather than a coincidental similarity. *Conf* is the confidence score assigned to the mapping, where confidence generally ranges from 0.0 to 10.0, *Source* is the source concept that maps onto the target concept of *cell*, and *Source Examples* and *Target Examples* are sentence fragments from the two corpora that evidence the mapping. Rows in the table can be read as "a cell is like a material, because creating from, mixing, and producing cells is like creating from, mixing with, and producing materials."

Note that the example sentences often use a variety of words that represent the concept involved in the mapping. For example, the source concept of *material* can be represented by many words: hydrogen, iron, fiber, water, air, element, solvent, and many others. CMI's use of WordNet enables it to find a wide variety of words that may represent the concepts involved in a metaphor. Similarly, while individual verbs such as make, create, or produce may not seem particularly metaphorical, it is the aggregate pattern of their repeated, systemic use that demonstrates the conceptual metaphor. Here again, we see the potential of CMI to foster critical thinking about metaphors. An individual phrase such as, "created from an already existing cell," might not, on its own, appear incredibly metaphorical. However, when seen as part of a larger pattern, these individual phrases can add up to suggest potential metaphorical framings.

Discussion

The top computationally identified metaphors shown here form three conceptual groups: things that are built, the raw materials from which things are built, and the ways in which things are built. The metaphors A CELL IS A BUILDING and A CELL IS A STRUCTURE fit into this second group and are indicative of the way that students refer to cells as things that are made, created, or produced. These metaphors also align to some extent with the conceptual metaphor that BODIES ARE BUILDINGS (Lakoff et al., 1991) , except that cells are the fundamental unit from which bodies are constructed, not bodies themselves.

Thus, it might appear that a different set of computationally identified metaphors—A CELL IS A MATERIAL, A CELL IS A PIECE, and A CELL IS WATER—aligns more closely with the BODIES ARE BUILDINGS metaphor, in that bodies are composed of cells, and buildings are composed of materials and pieces. However, while cells are often described as the basic unit of living organisms, they do not serve the role of raw materials in the body. Proteins or amino acids would likely be a better target for a metaphorical mapping from raw materials. Exposing this sort of potential misconception, both to teachers and to students, and enabling critical engagement with potential alternative metaphors is one of the potential strengths of CMI.

A third set of metaphors, that A CELL IS A STYLE and A CELL IS A DESIGN, help demonstrate CMI's potential to encourage creative generation of alternative metaphors. The previous two groups of metaphors have fit with the overall BODIES ARE BUILDINGS metaphor, in which cells are either constructed like buildings, or are the raw materials from which bodies are constructed. However, this later group of metaphors suggests a slightly different framing where in cells are a specific style or design pattern. This could be interpreted in at least two ways. First, an individual cell, and the DNA it contains, could be seen as a specific style or design for how to build an organism, such that, during cellular reproduction, one cell might "combine with" another to form a new, slightly different cell, along with a slightly different design or style for the organism of which the cell is a part. Second, since all living things are "made from" cells, the very notion of a cell could be seen as a style or design pattern from which all nature is constructed. Here, we see how CMI can generate results that are somewhat unexpected but still have the potential to promote critical or creative thinking about conceptual metaphors.

Conclusion

This paper presents the results of applying computational metaphor identification (CMI) to seventh grade students' writing about cellular reproduction. The results presented here demonstrate not only CMI's efficacy in identifying metaphors, but also its potential for fostering students' facility with noticing, critically examining, and creatively generating different metaphors. In future trials, the results from these and new analyses can be integrated into the WISE instructional module, allowing students to grapple with alternative metaphors and potential misconceptions. An important aspect of the results from this computational method is that multiple potential metaphors are identified for a single target concept. Lakoff and Johnson (1980, p. 221) describe the concept of metaphorical pluralism, that "successful functioning in our daily lives seems to require a constant shifting of metaphors... that are inconsistent with one another... to comprehend the details of our daily existence." As revealed by the same authors' later work, (1983), shifting metaphors may be necessary to effectively approach different aspects of the same concept. An important component of developing facility with metaphorical thinking is fostering awareness of metaphor. By computationally identifying potential metaphors in students' writing, this work seeks to foster critical thinking and creativity with metaphor in science learning.

References

- Blanchette, I., & Dunbar, K. (2001). Analogy use in Naturalistic settings: The influence of audience, emotion and goals. *Memory and Cognition*, 29, 730-735.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Gentner, D., & Gentner, D. (1983). Flowing waters or teeming crowds: Mental models of electricity. In *Mental Models* (pp. 99-129). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Gordon, W. (1974). Some source material in discovery-by-analogy. *Journal of Creative Behavior*, 8, 239-257.
- Kilgarriff, A. (2003, April 22). BNC Word Frequency List. http://www.kilgarriff.co.uk/bnc-readme.html.
- Klein, D., & Manning, C. (2003). Accurate Unlexicalized Parsing. In *Mtg of the Assoc for Comp Ling*. Sapporo, Japan.

- Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor* and thought, 2nd. ed. (pp. 202-251). New York: Cambridge Univ Press.
- Lakoff, G., Espenson, J., & Schwartz, A. (1991). *The Master Metaphor List*. Berkeley: Cognitive Linguistics Group, University of California, Berkeley.
- Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By* (2003rd ed.). Chicago, IL: University of Chicago Press.
- de Marneffe, M., MacCartney, B., & Manning, C. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Lang Res and Eval (LREC)*. Genoa, Italy. Retrieved from http://nlp.stanford.edu/pubs/LREC06_dependencies.pdf.
- Resnik, P. (1993). Selection and Information: A Class-Based Approach to Lexical Relationships. University of Pennsylvania, Department of Computer and Information Science.
- Zook, K., & Di Vesta, F. (1991). Instructional analogies and conceptual misrepresentations. *Journal of Educational Psychology*, *83*, 246-252.