Proceedings of DETC '04
ASME 2004 Design Engineering Technical Conferences and Computers and Information in Engineering Conference
September 28-October 2, 2004, Salt Lake City, Utah USA

DETC2004-57250

NATURAL LANGUAGE ANALYSIS FOR BIOMIMETIC DESIGN

I. Chiu and L.H. Shu*

Department of Mechanical and Industrial Engineering
University of Toronto
5 King's College Road, Toronto, Ontario M5S 3G8 Canada
*shu@mie.utoronto.ca

ABSTRACT

Biomimetic design uses ideas from biological phenomena as inspiration in design. To support biomimetic design, biological analogies are identified by finding instances of functional keywords that describe the engineering problem in biological knowledge in natural-language format. Challenges in using this approach include the identification of keywords, and the quantity and quality of results found.

WordNet, a lexical database, is used as a language framework to systematically generate alternative keywords to find matches and analyze the results of searches. Troponyms from WordNet were found to provide better and more plentiful keywords than did synonyms.

Due to the potentially large number of matches to keywords, matches are analyzed to facilitate extraction of dominant biological phenomena associated with keywords. This analysis found that words that frequently collocated with keywords tend to be objects of the keyword verb or agents that carry out the actions of the keyword. Furthermore, nouns that are inanimate, e.g., substances, tend to be objects, and nouns that are animate e.g., animals, organs, tend to be agents. Distinguishing frequently collocated words and their relationships to keywords can be used to facilitate identification of biological analogies in natural-language format to support design.

1 INTRODUCTION

Biomimetic design uses ideas from biological phenomena to inspire design concepts. Many examples of biomimetic design exist, the most often cited of which is the development of Velcro after observation of how plant burrs stick to materials such as clothing and fur. Most instances of biomimetic design occur following observation of an interesting biological phenomenon or were inspired from a biological phenomenon already known to the designer. We believe that a systematic search of biological phenomena relevant to a particular design problem will identify a greater variety of potential analogies, and likely result in more creative design than depending on chance knowledge of biological phenomena.

One approach to support a systematic biomimetic design process would be to create a database of biological phenomena relevant to engineering (Vincent and Mann, 2002). However, not only is the task of creating and updating such an immense database significant, but the assignment of biological phenomena to engineering categories is a subjective process that may reduce the richness of the original information.

Our approach to support biomimetic design is to directly search the vast amount of biological knowledge already available in natural-language format. Previous work using this approach, described by Vakili and Shu (2001), Hacco and Shu (2002), and Shu et al. (2003) searches for instances of keywords and their synonyms in a biology text, *Life, the Science of Biology* (Purves et al., 2001).

A search to support biomimetic design imposes additional challenges beyond simply searching by topic, since the relevance of potentially analogous information across disciplines, or between domains, is more difficult to determine than the relevance of direct information on a specific topic.

2 MOTIVATION

Several difficulties common to natural-language processing occurred during previous work, from the identification of suitable keywords, to the quality and quantity of matches that result. These difficulties serve as motivation for the work reported in this paper.

2.1 Identification of appropriate keywords

Keywords that describe the engineering problem may be used neither in a biological context nor in everyday speech. Therefore, alternative keywords are required to increase the chance of finding relevant results. Previous work used synonyms as additional keywords. However, the relationship between the original keyword and synonyms identified from a thesaurus is not always clear.

2.2 Quality of matches

Previous work that searched for occurrences of keywords and their synonyms in the text resulted in several irrelevant matches in addition to the relevant matches (Hacco and Shu, 2002). One factor that contributed to the difficulty was that the part-of-speech of the search keyword versus occurrences of the keyword found were not taken into account. The ability to distinguish between potentially relevant and definitely irrelevant results is critical to the utility of a search tool.

2.3 Quantity of matches

Although the initial body of knowledge searched is limited to a single textbook, the number of matches to keywords and their synonyms can become unmanageable. The task of reviewing all the matches can be tedious and time-consuming, during which relevant matches can be overlooked. In addition, the relevance of matches that identify unfamiliar biological phenomena cannot often be determined until further understanding of such phenomena is obtained. Therefore, a summarizing mechanism is required to quickly point the designer to the most promising results, and to target further research.

Benami and Jin (2002) support that novel stimuli to enhance creativity must be relevant and meaningful; otherwise designers will waste time analyzing such information without producing creative ideas.

This work addresses the identification of suitable keywords to use in searching for relevant biological phenomena, which determines the quality of matches found. Also addressed is how to reduce the amount of information that the designer must review after a search to identify relevant information.

A tool that is central to this work is WordNet, a lexical database whose organization is based on how words are believed to be stored in human memory rather than in an alphabetized list. We first define terminology from the field of linguistics, WordNet, as well as other terminology frequently used in the paper.

3 NOMENCLATURE

Collocation – The occurrence of a word in association with another word, usually the keyword used for searching. Also referred to as a co-occurrence.

Corpus – A written sample of language usage for linguistic analysis.

Function Words – Words belonging to grammatical or function classes such as articles, conjunctions and prepositions (Akmajian et al, 1998).

Hypernym – Describes the superset of a word, where the hypernym encompasses all instances of x. For example, tree is the hypernym of maple (Miller, 1993).

Hyponym – Describes the subset of a word, where the hyponym is a specific instance of y. For example, maple is a hyponym of tree; tree is a hyponym of plant (Miller, 1993).

Keywords – Used to search for text documents or passages that contain instances of these words.

Keyword-match passage – Text segment that contains the sought keyword.

Sense – The meaning of a word. Words may have multiple senses or meanings. Senses in WordNet are enumerated.

Troponym – Specifically refers to the hyponym relationship between verbs. The relationship between two verbs is V1 is to V2 in some particular manner (Fellenbaum, 1993). For example, "to amble" is a troponym of "to walk" because ambling is a particular manner of walking.

4 METHODS

The approach taken is based on ideas that have been presented by the natural language processing and computational linguistics community, particularly latent semantic indexing (Deerwester et al., 1990, Yu et al., 2003) and word collocations (Yarowsky, 1995).

Our methods will be illustrated by an example: seeking biological analogies for "cleaning", e.g., dirt from clothes.

4.1 Use of WordNet

A combination of human and computational processing is currently used. Future work includes methods for further automation. As a language framework, WordNet (Miller, 1993) is used. WordNet is an electronic lexical database, designed and organized according to current psycholinguistic and computational theories of how humans remember language. This is very different from a dictionary or thesaurus in that word entries are not alphabetical but rather, they are organized based on the relationships to other words.

Two features of WordNet that were particularly useful for this approach are the word sense categories and the troponym trees. Word senses refer to the different meanings of a word. For example, querying the verb "remove" within WordNet, the first sense of "remove" is "[to] remove something concrete as by lifting, pushing, taking off, etc." The second sense of "remove" is "[to] remove from a position of an office". For the verb "remove", there are a total of eight senses or different meanings. Based on the word senses given within WordNet, it is possible to sort biomimetic search results by the meaning of the word when the context of the word is considered.

Troponyms can also be thought of as the hyponym or "subset" of verbs that describe specific manners of another verb, i.e., a troponym is a "particular way of doing X". For the first sense of the verb "remove", a troponym is "skim," which is applicable to the removal of a layer from the surface of liquids.

4.1.1 Step 1: Identify Keywords

As in our past work, this work uses verbs as the keywords for searching. Stone and Wood (1999) used verbs to convey function in their functional basis for design. McAdams and Wood's work on design-by-analogy (2000) discusses how objects can be very different in form but still share a functional commonality. Some researchers use "function" interchangeably with "intended behavior" (Shooter et al., 2000). Benami and Jin (2002) found that stimuli presented as forms and behaviors are more effective than stimuli presented as functions. We embody desired functions into verbs, and search these verbs to locate biological forms and behaviors to be used as stimuli.

Furthermore, a biomimetic search that uses verbs rather than nouns will less likely bias the designer towards a possibly preconceived biological phenomenon, but instead introduce new phenomena and thus provide new analogies. For example, searching for "kidneys" which are known to remove toxins from blood will only provide matches with "kidneys" while searching for "remove" will provide matches with other subjects that "remove" as well as objects that are removed.

In our past work, synonyms to keywords were used to increase the number of matches. Internet search engines also use synonyms to increase the number of matches. In this work, a combination of hypernyms and troponyms is chosen over synonyms to provide alternative keywords. Troponyms were found to provide better alternative keywords than synonyms from a thesaurus as well as the synonym feature within WordNet. In addition, using troponyms made it possible to work within the same word sense structure in WordNet. Finally, the troponym feature in WordNet provided an extensive list of other actions not necessarily included in a synonym list. For example, the troponym listing for the first sense of "remove" provided 179 results, while the synonym listing for the first sense of "remove" only provided 4 results, 3 of which are contained in the troponym list. While it is difficult to judge the quality of the keywords generated using either troponyms or synonyms, the use of more keywords will not only provide a greater quantity of matches, but also a more varied, richer, assortment of matches.

For our cleaning example, the verb "clean" is an obvious keyword with which to start. The troponym list for "clean" includes many human-specific methods of cleaning e.g., vacuuming, flossing and soaping. Instances of these words in the Purves et al. (2001) text correspond only to the root noun forms, e.g., references to the speed of light in a vacuum and cells that have a similar shape to soap bubbles. Therefore searching this text for the troponyms of "clean" is not likely to identify biological analogies for cleaning.

However, since "clean" is a troponym, i.e., specific instance of "remove," the word "remove" may be used to find analogies for "clean". The troponyms of the verb "remove" consist of 179 verbs that represent different methods of removing.

4.1.2 Step 2: Search corpus for keywords

Searching the Purves et al. (2001) text for instances of the 179 troponyms for "remove" resulted in matches for 38 of them, including the word "remove" itself, listed in Table 1.

Keyword-match passages containing these troponyms were stored as plain text files as well as entered into a spreadsheet.

4.1.3 Step 3: Remove unlikely relevant matches

Removal of matches unlikely to be relevant to the cleaning problem was performed to reduce the number of keyword-match passages. Several factors were observed to result in irrelevant matches.

Table 1: Troponyms for "remove" with matches

1. Abrade	2. Abscise	3. Amputate
4. Bail	5. Break out	6. Brush
7. Circumcise	8. Clean	9. Clear
10. Delete	11. Detoxify	12. Dig
13. Discharge	14. Dislodge	15. Divert
16. Draw	17. Eliminate	18. Empty
19. Excavate	20. Excrete	21. Harvest
22. Kill	23. Leach	24. Pull
25. Pump	26. Rasp	27. Remove
28. Scavenge	29. Scoop	30. Shed
31. Skim	32. Stem	33. Strip
34. Suck	35. Tap	36. Unload
37. Wash	38. Withdraw	

a. Non-verb instances of keywords

Matches to keywords where the keyword was used in a part of speech that is not a verb were removed. This was performed because the noun form of a word often has a significantly different meaning from the verb form. For example, the verb "to strip" means to "deprive, divest" and the noun "strip" means "a relatively long narrow piece of something" (WordNet 2.0). Thus, matches to the noun sense of strip were removed from the results. Although some noun forms are related to the verb form of a word, e.g., for the word "pump", currently matches with all nonverb forms of the keyword are removed.

b. Keywords acting on "abstract" objects

In this example involving cleaning, we are interested in analogies that involve the removal of physical, as opposed to, abstract objects. Therefore, instances where the verb acts on an abstract entity are removed from the search results. For example, for the troponym "eliminate," matches referring to "eliminating risk" were removed.

The removal of abstract verb-object pairings was also performed when finding biological analogies for a problem in design for repair and remanufacture (Hacco and Shu, 2002).

c. Keyword verbs used in different sense

Many verbs are polysemous, i.e., they have multiple meanings. Although the search words are troponyms of "remove", many have additional meanings that are not related to "remove". For example, the word "draw" can be used in the sense, "to draw out," or remove, water, or it can mean, "to make or trace a figure" as in to draw a diagram. While methods to draw water may be relevant to the process of removing or cleaning, methods to draw a picture or a figure are not. Therefore, matches where these search words are used in a sense unrelated to "remove" were discarded.

4.1.4 Step 4: Find frequently co-occuring words

The goal of this step is to quantitatively determine dominant biological phenomena by identifying the words that most frequently collocate with, or occur in passages of text that contain, the keywords. Specifically, words that occur within a 50-word window around the keyword were counted and sorted.

To reduce noise introduced by frequently occurring function words such as "the", "a", "is", a stop list was used to exclude these words from being counted. The stop list included the most frequently occurring words in the English language (Yu, et al., 2003) as well as other function words observed to occur frequently in the matches. Other additions to the stop list include the spelled-out forms of numbers and single letters of the alphabet in the context of "diagram a, b, c, etc." Such a stop list may require adjustments specific to the corpus searched. The keyword itself was also excluded from being counted. For regular verbs, instances of the keyword verb in other tenses, e.g., past tense, were also excluded.

The plain text files generated in Step 2 that contain the keyword matches were used as input to a script. The output files contain all words (except those excluded above) that occurred in the input file and the number of times they occurred, sorted in descending frequency of occurrence. Also included was the section number of the corpus, i.e., Purves et al. (2001), where each match was found.

Figure 1 contains a section of the output frequency file for "remove." The word "cells" occurs the most frequently in the vicinity of the word "remove", a total of 31 times, the first two instances of which were in section 17.2.3 of the text.

	A	B	C	D	E
129	**Word**	**Frequency**	**Section**		
130	cells	31	17.2.3	17.2.3	17.5.2
131	dna	24	14.5.3	14.5.3	17.2.3
132	figure	17	12.9.2	14.7.2	19.6.2
133	species	17	33.2.2	33.2.2	54.6.2
134	cell	16	12.9.2	17.2.3	19.6.2
135	other	15	19.2.1	21.4.1	33.2.2
136	blood	15	17.5.2	17.5.2	41.4.4
137	protein	15	12.8.1	14.4.2	3.3.8
138	because	14	17.5.2	33.2.2	50.6.2

Figure 1: Word frequency file for keyword "remove."

4.1.5 Step 5: Analyze results

The word frequency file with fixed-width columns was used to examine results visually without graphing. For example, Figure 2 shows a portion of the same file as shown in Figure 1, but at a much smaller scale. Similarly, Figure 3 shows a portion of the minimized spreadsheet containing the word frequency counts for "eliminate". Files containing a few hundred unique words correspond to the presence of several matches for the keyword, since text passages of 50 words around matched keywords are analyzed.

The patterns in frequency counts shown in Figures 2 and 3 are similar for the two keywords. Word frequency files that correspond to larger numbers of matches exhibit a very steep drop-off in the word count, with most of the unique words occurring once. For keywords where there were fewer than three keyword-match passages in the entire text, this pattern does not appear to hold as strongly.

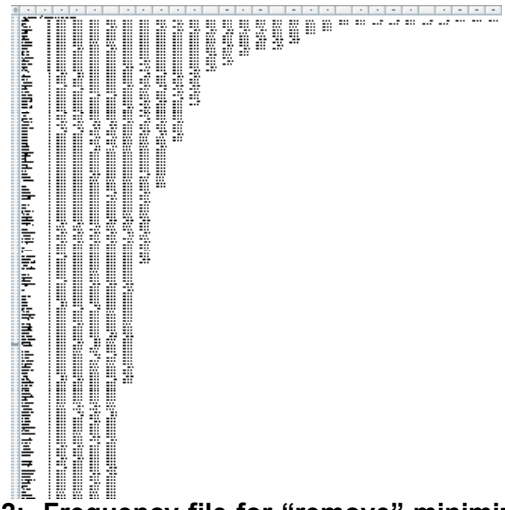

Figure 2: Frequency file for "remove" minimized.

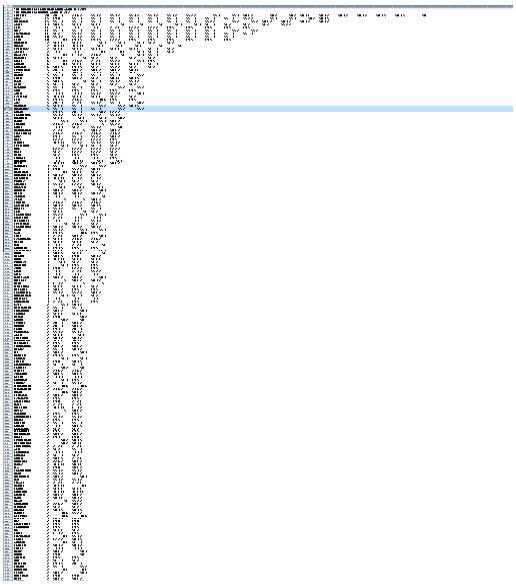

Figure 3: Frequency file for "eliminate" minimized.

a. Frequency cut-off

Determining a frequency cut-off to distinguish significant collocated words would facilitate analysis of the collocation frequency files, since some keywords have more than 1200 unique word collocations, most of which are single occurrences. The frequency density curves were observed to approximate a chi-squared distribution with one degree of freedom, and a target critical value of 0.05 was used to define which collocated words were significant.

This critical value corresponds to approximately the top 5% of the unique word occurrences with respect to frequency. The actual cut-off selection depends on factors such as the total number of matches for the keyword, and whether a particular collocated word directly corresponds to the 0.05 cut-off. Since the actual distributions are not ideal, the actual cut-off ranged between 0.038 and 0.084. Such cut-offs may occur when there are large gaps between the frequencies of words.

b. Dominant theme associated with keyword

Examining the most frequently collocated words can be used to capture the dominant biological theme that is associated with each of the troponym keywords. For example, the troponym "harvest" is associated with energy (most frequently collocated noun with 18 instances) while the troponym verbs "bail" and "excrete" are associated with water (5 and 57 occurrences respectively).

Further insight into the dominant biological themes associated with keywords is gained by exploring the relationships between keywords and frequently collocated words. To this end, keyword-match passages from Step 2 were examined to determine how the most frequently collocated words were used.

Examination of the keyword-match passages revealed that the frequently collocated words were likely to be the direct object (to be referred to as "object" henceforth) of the keyword, i.e., what the verb is acting on. Therefore, the list of the most frequently collocated words for the remove/clean example will often answer the question "what is being removed?" Less often, the frequently collocated word appears as the subject, or agent performing the action of the keyword.

For each passage, the agent and object of the keyword, i.e., the agent that performs the keyword action, and the object that the keyword acts on, were identified. It is realized that "agent" and "object" are not used in the strictest sense of English grammar. The agent can be considered the "doer" and the object, the "doee". For example, in a match for "excrete", "Nasal salt glands excrete excess salt," the agent doing the excreting is "glands" and the object being excreted is "salt".

The agent/object identification is only performed within the sentence. When the agent or object is named outside the sentence, and replaced in the sentence by a pronoun, a language phenomenon called anaphora (Akmajian et al., 1998), the agent/object was not identified. When the agent/object is an entire noun phrase, only the noun itself was identified as either the agent or object. For example, in the above noun phrase "nasal salt glands", the identified agent was "glands".

While only the keyword and agent/object relationships have been classified with respect to frequency, there are other relationships that merit examination. In addition to noun phrases, including the use of nouns as adjectives, e.g., "salt glands", indirect objects and prepositional phrases can also be investigated for their contribution to meaningful words in the frequency list.

Figure 4 shows the overlap between agents/objects from the keyword-match passages of Step 2 and the most frequently collocated words of Step 4. These overlapped words are used to determine the dominant biological phenomena associated with a particular keyword. In the case of frequently occurring objects, this information can help target further investigation on how an object is removed and even why the object is being removed, thus providing the designer with potential biological analogies without examining all the matches.

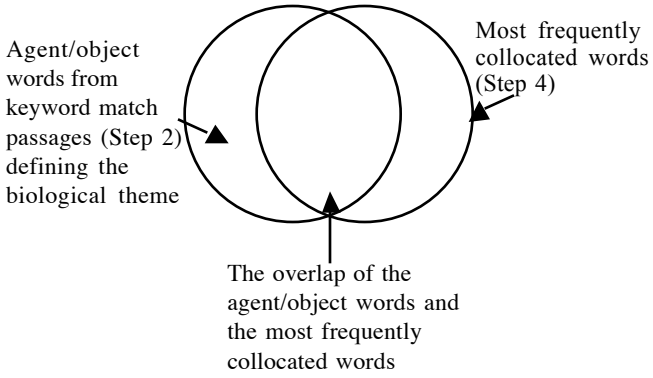

Figure 4: Dominant biological phenomena in overlap between agents/objects and frequent collocations.

5 EXAMPLE TROPONYM: ELIMINATE

The above process is illustrated for the troponym "eliminate". Table 2 shows for "eliminate" up to the cut-off frequency of 4, the collocated words, the number of occurrences, and the number of times the collocated word acted as an agent or was the object. This cutoff frequency corresponds to the top 4% of all collocated words. One possible analogy to cleaning or removal suggested by the information in Table 2 concerns "elimination of species by predators." This is only one potential analogy identified by examining only 4% of the text contained in corresponding matches.

Table 2: Comparison of word frequencies and word usages in keyword matches for "eliminate."

Count	Word	Object	Agent	Count	Word	Object	Agent
20	Species	7	2	5	Nymphs	-	1
15	Predators	-	3	5	Mhc	-	-
13	Water	2	-	5	Systems	-	1
11	Cell	1	-	5	Habitats	1	-
11	Prey	3	-	5	Dragonfly	-	-
11	Environments	-	-	4	Competition	-	-
10	Cells	2	-	4	Products	2	-
7	Nitrogen	1	-	4	Range	-	-
7	Blood	1	-	4	Change	-	-
7	Excretory	-	-	4	Population	1	-
7	Wastes	3	-	4	Classifications	-	-
6	Balance	-	-	4	Host	-	-
6	Organs	-	1	4	Reduce	-	-
6	Animals	-	-	4	Excreting	-	-
5	Extinction	-	1	4	DNA	-	-
5	Ponds	-	-	4	Base	-	-
5	Food	-	-	4	Help	-	-
5	Acid	1	-	4	Surface	-	-

Table 3 shows the percentage of biological themes covered within the most frequently collocated words that act as object or agent for the keyword, to all the individual keyword matches from the textbook. The percentage was calculated as follows:

1. For a given keyword's matches, all unique objects or agents were noted along with the frequency at which these objects or agents occurred.

2. From the frequency count of words collocated with the keyword, all words until the cut-off were examined and occurrences of these words as objects or agents were noted.

3. Finally calculated is the ratio of the number of matches containing either a frequently collocated object or agent to the total number of keyword-match passages. This metric is meant to approximate the percent coverage by frequent agents and objects of all phenomena presented in keyword-match passages.

Table 3: Percent of keyword matches covered by the most frequently collocated words

Troponym keyword	Number of keyword matches	Cut-off frequency	All occurrences of frequent objects	Percent of matches covered by frequently occurring objects	All occurrences of frequent agents	Percent of matches covered by frequently occurring agents	Percent of matches covered by frequently occurring objects and agents
Draw	19	3	13	68.4%	3	15.8%	68.4%
Eliminate	45	4	25	55.6%	9	20.0%	60.0%
Excrete	58	8	46	79.3%	23	39.7%	89.7%
Harvest	22	2	16	72.7%	6	27.2%	81.8%
Kill	91	5	66	72.5%	30	33.0%	75.8%
Pull	44	4	23	52.3%	16	36.4%	68.2%
Pump	45	5	41	91.1%	20	44.4%	93.3%
Remove	125	5	60	48.0%	19	15.2%	53.6%
Shed	18	3	12	66.6%	5	27.8%	72.2%

The first two percentages in Table 3 may sum to greater than 100 since frequent object collocations and frequent agent collocations were counted independently. The last column is calculated by counting only once instances where both the agent and object of the keyword are frequent. The last column shows the proportion of keyword matches covered by the most frequently collocated agents and objects, and is calculated as:

[# of frequent objects + # of frequent agents - # frequent object and frequent agent in same match] / # matches x 100%. A higher coverage of phenomena can be achieved by increasing the number of words included within the cut-off.

5.1 Objects of "eliminate"

Continuing with our example, the troponym keyword "eliminate" had 45 total matches. Table 4 shows the frequency of the words that acted as objects within keyword-match passages for "eliminate." The sum of the unique object frequencies in Table 4 is one greater than 45 due to one instance of direct objects, "water and carbon dioxide" where each object contributed separately to the frequency count.

Table 4: Complete set of objects from keyword-match passages for "eliminate"

Object	Number of occurrences as object of keyword	Frequently collocated w/ keyword?
Acid	1	Yes
Blood	1	Yes
Cell, cells	3	Yes
Chestnut	1	
Competitor	1	
Dioxide	1	
Gene, genes	2	
Groups	1	
Habitats	1	Yes
HIV	1	
Malaria	1	
Matter	1	
Molecule	1	
Monkeys	1	
Mussels	1	
Nitrogen	1	Yes
Population	1	Yes
Prey	3	Yes
Products	2	Yes
Pseudogenes	1	
Recombination	1	
Shrubs	1	
Smallpox	1	
Species	7	Yes
Subpopulation	1	
Tadpoles	2	
Taxa	1	
Vitamins	1	
Wastes	3	Yes
Water	2	Yes

From the frequency counts of words collocated with "eliminate" shown in Table 2, 11 of the object words in Table 4 fall within the cut-off defining most frequently collocated words: acid, blood, cell(s), habitats, nitrogen, population, prey, products, species, wastes and water. The cut-off used for Table 2 was 4 occurrences or more, determined using the 0.05 critical value. In other words, about 95% of the words associated with "eliminate" occurred three times or fewer.

From this, the percentage of matches found within the most frequently collocated objects is calculated as:

[acid x 1 + blood x 1 + cell(s) x 3 + habitats x 1 + nitrogen x 1 + population x 1 + prey x 3 + products x 2+ species x 7 + wastes x 3 + water x 2] = 25 / 45 x 100% = 55.6%.

Therefore, more than half of the biological phenomena (eliminating acid, blood, cells, habitats, nitrogen, population, prey, products, species, wastes and water) associated with the

keyword "eliminate" can be gleaned by simply examining the relationship between the keyword and the above 11 words. These 11 objects, selected from the top 4% of unique words collocated with "eliminate", represent about 1.5% of the unique collocated words, examination of which is a significant reduction in effort from reading all 45 keyword-match passages.

As is the case with many keywords, matches with "eliminate" identified phenomena from multiple levels of biological organization, from cell-level concepts such as gas exchange to community-level concepts such as the interaction of species. These phenomena are covered by the intersection of the most frequently collocated words and objects of the keyword. The ability to find phenomena from multiple levels, and thus physical scales, provides a richer diversity of potential analogies than those that naturally "come to mind", which typically range from the organ to organism levels.

5.2 Agents of "eliminate"

Although frequently collocated words tend to be objects of the keyword, they can also appear as agents that perform the keyword function. Table 2 showed that for "eliminate", 25 instances of frequently occurring words were used as objects and 9 were used as agents. This imbalance between objects and agents can be due to the typical use of the passive voice in scientific writing, where the agent is missing from the sentence altogether. For example, in "When the enclosures were removed, arthropods quickly recolonized the islands," the question of "what did the removing?" is not explicitly answered in the sentence. However, in other cases, such as the previously used example, "Nasal salt glands excrete excess salt", the agent is clearly identified. For such sentences, in addition to the keyword-object relationship, the agent-keyword relationship can be identified. Therefore, biological phenomena associated with a keyword can also be obtained by examining the agent. Table 5 shows the complete set of agents for "eliminate," examination of which reveals predators as one agent of elimination in biological phenomena.

5.3 Distinguishing between agents and objects

The WordNet lexical hierarchy partitions nouns into hierarchies corresponding to semantic fields with distinct concepts (Miller, 1998). Nouns are classified into lexical categories called unique beginners, examples of which include event, substance and body, which are then further classified. Words classified under substance in the WordNet hierarchy tend to be objects. For examples from Table 2, "water" is a compound-type substance, while "nitrogen" is a chemical element-type substance (WordNet 2.0), both of which are inanimate. From the agents column, "predators" and "nymphs" are classified as animals and animate beings while "organs" are body part nouns. Both animal and body part nouns tend to have more animate roles and the ability to act upon something. An interesting agent from Table 2 is "extinction" which is classified as a process noun or an event noun, also with the ability to act on something.

However, WordNet noun categories cannot always be used to distinguish between agents and objects. For example, the word "species" is usually collocated with "eliminate" and other troponyms as an object, but it can also occur as an agent. While species is a group-type noun with some animate possibilities, "water", an inanimate object, can also be the agent. For example, when collocated with the troponym "leach", water is the agent that leaches from seeds inhibitors that prevent their germination.

One interesting finding is that it is very rare for words that are not nouns to occur frequently. In the specific example for "eliminate", only one adjective, "excretory", appears within the most frequently occurring words. Only four verbs, "change", "reduce", "excreting," "help" occur frequently, and correspond to the lowest frequencies included within the cut-off.

Table 5: Agents within matches for "eliminate"

Agent	Agent frequency in passages	Frequently occurring word?
Area	1	
Blight	1	
Death	1	
Deforestation	1	
Deletion	1	
Destruction	1	
Diffusion	1	
Disturbances	2	
Drugs	1	
Expansion	1	
Extinction	1	Yes
Insect	1	
It	2	
Kidney	1	
Leeches	1	
Lysis	1	
Nymphs	1	Yes
Organization	1	
Organs	1	Yes
Predators	3	Yes
Pressure	1	
Rats	1	
Scour	1	
Selection	1	
Slumping	1	
Species	2	Yes
Stars	1	
Systematists	1	
Systems	1	Yes
They	1	
Unit	1	

5.4 Application of dominant biological theme

Examination of the frequent objects and agents for the keyword "eliminate" reveals several biological themes, one of which is species interaction, e.g., through competition and predator and prey relationships. The species interaction phenomenon involves competition for resources and occurs at the ecology level of biology. Applying the idea of competition for resources to the cleaning problem, one possible solution is to use a less porous surface that provides fewer locations (resources) for dirt to settle. Another possible solution is to reduce dirt's access to the surface (resource) through the presence of competing substances, such as protective coatings and sprays. In both solutions, the idea of reducing the resources available to dirt is applied to the cleaning problem. Mak and Shu (2004) further describe the extraction of strategies from descriptions of biological phenomena and application of these strategies to an engineering problem.

6 NATURAL LANGUAGE ANALYSIS AND DESIGN

The reduction of biological knowledge in natural-language format to verb-object and agent-verb relationships greatly simplifies the identification of possible biological analogies for a given design problem. The verb-object and agent-verb relationships follow from English-language grammar rules. These natural-language rules are simple enough that they can be exploited to succinctly summarize a relatively large amount of biological knowledge for use in design. Much work has been done to define grammars for design purposes specifically because of the ability to methodically apply simple rules for design (Li and Schmidt, 2000) and for design synthesis (Starling and Shea, 2002). Using a natural-language corpus to support our design process, simple language-grammar based rules are formulated to analyze text. These rules are further supported by the use of WordNet as a language framework.

Other applications of language analysis in design include the capture of design knowledge through noun phrase extraction and mapping (Yen, Fruchter and Leifer, 1999). Our focus is on verbs that express function, further supplemented by noun, object or agent, identification to provide a more complete view of potential analogies.

7 SUMMARY

A natural-language approach to support biomimetic design was chosen to avoid the immense task of categorizing all of biological phenomena for engineering purposes, and to take advantage of the enormous amount of biological knowledge already in natural-language format. Previous work using this approach involved difficulties that include the identification of keywords to search for analogous biological phenomena, and the quantity and quality of results found. WordNet, a lexical database that contains senses and troponyms for words, is used as a language framework to systematically generate alternative keywords, find matches and analyze the results of searches.

In previous work, synonyms of keywords describing desired engineering function were used to increase the number of relevant matches in a biological corpus. In the current work, troponyms were found to provide better and more plentiful additional keywords than did synonyms.

In previous work, a large quantity of matches, including both relevant and irrelevant results, decreased the practical value of the search method. In the current work, distinguishing the words that frequently collocate with keywords and their relationships with the keyword reduces the effort required to identify dominant biological phenomena. Frequencies of words that collocate with keywords were generated and exhibit a common pattern when several matches were found for the keyword. Specifically, sharp drop-offs in the collocated word frequencies led to the definition of frequency cut-offs to distinguish significant and non-significant collocations.

Frequently collocated words tend to be objects of the keyword verb or agents that carry out actions of the keyword. Furthermore, nouns that are inanimate, e.g., substances, tend to be objects, and nouns that are animate e.g., animals, organs, tend to be agents. Distinction between animate and inanimate objects can be performed using WordNet.

There were more objects than agents due to the wide use of the passive voice in scientific writing. Therefore, dominant biological themes can often be described by the keyword-object relationship. This result corresponds well with Stone and Wood's (1999) use of verb-object pairings in their functional basis for design.

This work supports creativity in design by identifying biological phenomena relevant to any given design problem through natural-language analysis of biological knowledge. A method to significantly reduce the effort required to systematically identify relevant biological phenomena from natural-language knowledge was presented.

8 ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support of NSERC (Natural Sciences and Engineering Research Council of Canada) through a Discovery Grant.

9 REFERENCES

Akmajian, A., Demers, R. A., Farmer, A. K. and Harnish, R. M., 1998, *Linguistics, An Introduction to Language and Communication, Fourth Edition*. MIT Press, Cambridge, MA.

Benami, O., Jin, Y., 2002, Creative Stimulation In Conceptual Design, *Proceedings of ASME DETC/CIE*, Montreal, QC, Canada, DETC2002/DTM-34023.

Deerwester, S., Dumair, S. T., Harshman, R., 1990, Indexing by Latent Semantic Analysis, *Journal of the Society for Information Science, 41(6)* pp. 391-407.

Fellenbaum, C., 1993, English Verbs as a Semantic Net, *Five Papers on WordNet,* pp 40-61.
ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps

Hacco, E., Shu, L.H., 2002, Biomimetic Concept Generation Applied to Design for Remanufacture, *Proceedings of ASME DETC/CIE,* Montreal, QC, Canada, DETC2002/DFM-34177.

Li, X., Schmidt, L., 2000, Grammar-Based Designer Assistance For Epicyclic Gear Trains, *Proceedings of ASME DETC/CIE*, Baltimore, MA, DETC2000/DTM-14574.

Mak, T, Shu, L., 2004, Use of Biological Phenomena in Design by Analogy, *Proceedings of ASME DETC/CIE,* Salt Lake City, UT, DETC2004/DTM57303.

McAdams, D., Wood, K., 2000, Quantitative Measures For Design by Analogy, *Proceedings of ASME DETC/CIE*, Baltimore, Maryland, DETC2000/DTM-14562.

Miller, G.A., 1993, Introduction to WordNet: An On-line Lexical Database, *Five Papers on WordNet,* pp. 1-25. ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps

Miller, G.A., 1998, Nouns in WordNet, *WordNet, An Electronic Lexical Database,* MIT Press, Cambridge, MA, pp. 23-46.

Purves W.K., Sadava, D., Orians, G.H., Heller, H.C., 2001, *Life, The Science of Biology*, 6/e, Sinauer Associates, Sunderland, MA.

Shooter, S., Keirouz, W., Szykman, S., Fenves, S., 2000, A Model For Information Flow In Design, *Proceedings of ASME DETC/CIE*, Baltimore, Maryland, DETC2000/DTM-14550.

Shu, L., Lenau, T., Hansen, H., Alting, L., 2003, Biomimetics Applied to Centering in Microassembly, *Annals of the CIRP*, 52/1:101-104.

Starling, A., Shea, K., 2002, A Clock Grammar: The Use of A Parallel Grammar in Performance-Based Mechanical Synthesis, *Proceedings of ASME DETC/CIE*, Montreal, QC, Canada DETC2002/DTM-34026.

Stone, R.B., Wood, K.L., 1999, Development of a Functional Basis for Design, *Proceedings of ASME DETC/CIE*, Las Vegas, NV, DETC99/DTM-8765.

Vakili, V., Shu, L.H., 2001, Towards Biomimetic Concept Generation, *Proceedings of ASME DETC/CIE*, Pittsburg, PA. DETC2001/DTM-21715.

Vincent, J., Mann, D., 2002, Systematic Technology Transfer from Biology to Engineering, *Philosophical Transactions of The Royal Society: Physical Sciences*, 360:159-173.

WordNet 2.0, http://www.cogsci.princeton.edu/~wn/

Yarowsky, D., 1995, Unsupervised Word-sense Disambiguation Rivalling Supervised Methods, *Proceedings of 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189-196.

Yen, S., Fruchter, R., Leifer, L., 1999, Facilitating Tacit Knowledge Capture and Reuse in Conceptual Design Activities, *Proceedings of ASME DETC/CIE*, Las Vegas, NV, DETC99/DTM-8781.

Yu, C., Cuadrudo, J., Ceglowski, M., Payne, J, 2003, Patterns in Unstructured Data: Discovery, Aggregation and Visualization, *Presentation to Andrew W. Mellon Foundation,* http://javelina.cet.middlebury.edu/lsa/out/lsa_explanation.