Jane Morris & Clare Beghtol
Faculty of Information Studies, University of Toronto

Graeme Hirst
Department of Computer Science, University of Toronto

Term relationships and their contribution to text semantics and information literacy through lexical cohesion

Abstract: An analysis of linguistic approaches to determining the lexical cohesion in text reveals differences in the types of lexical semantic relations (term relationships) that contribute to the continuity of lexical meaning in the text. Differences were also found in how these lexical relations join words together, sometimes with grammatical relations, to form larger groups of related words that sometimes exhibit a more tightly-knit internal structure than a simple chain of words. Further analysis of the lexical semantic relations indicates a specific need to focus on a neglected group of relations, referred to as *non-classical relations*, and a general need to focus on relations in the context of text. Experiments with human readers of text are suggested to investigate these issues, as well as address the lack of research that uses human subjects to identify reader-oriented relations. Because lexical cohesion contributes to the semantic understanding of text, these reader-oriented relations have potential relevance to improving access to text-based information. As well, the structured groups of words formed using a combination of lexical and grammatical relations has potential computational benefits to lexical cohesion analysis of text.

Résumé: Une analyse des approches linguistiques pour déterminer la cohésion lexicale d'un texte révèle les différences entre les types de relations sémantiques lexicales (relations terminologiques) contribuant à l'uniformité de la signification lexicale d'un texte. Des différences furent également identifiées dans la manière dont ces relations lexicales unissent les mots, parfois à l'aide de relations grammaticales pour former de plus grands ensembles de mots reliés démontrant parfois une structure interne plus étroitement liée qu'une simple chaîne de mots. Des analyses additionnelles des relations sémantiques lexicales indiquent un besoin spécifique de se concentrer sur une catégorie de relations souvent négligées, soit les relations dites *non-classiques,* de même qu'un besoin général de porter une attention particulière sur ces relations à partir du contexte textuel. On suggère de faire des expériences avec des lecteurs de texte afin d'étudier ces questions, de même que pour souligner le peu de recherches entreprises avec l'aide de sujets humains pour identifier les relations basées sur le lecteur. Étant donné que la cohésion lexicale contribue directement à la compréhension sémantique du texte, ces relations orientées vers le lecteur montrent la pertinence nécessaire pour améliorer potentiellement l'accès à l'information textuelle. De plus, les ensembles structurés de mots formés à la fois par l'utilisation des relations lexicales et grammaticales procurent un avantage computationnel à l'analyse de la cohésion lexicale du texte.

1. INTRODUCTION

When people read text, relations between words contribute to their understanding of it. While this is obvious, explaining exactly how is not. Otherwise computers would be much better at "understanding" (i.e., producing programmed analysis of) text semantics than they are. This paper describes an investigation into one aspect of this complex problem: the lexical semantic relations (term relationships) that create lexical cohesion in text (Halliday & Hasan, 1976).

Lexical cohesion occurs when related word pairs join together to form larger groups of related words that can extend freely over sentence boundaries. These larger word groups contribute to the meaning of the text through "the *cohesive* effect achieved by the continuity of lexical meaning" (Halliday & Hasan, 1976, p. 320, emphasis added). *Cohesion* is the general term for the linguistic features present in text that have been identified as "contributing to its total unity" (1976, p. 2). It "shows how sentences, which are structurally independent of one another, may be linked together through particular features of their interpretation" (1976, p. 10)[1]. These particular features are those that "occur where the *interpretation* of some element in the discourse is dependent on that of another", and hence are semantic in nature (1976, p. 4, original emphasis). In lexical cohesion, the dependency is simply that there be a recognizable relation between two words; that is, a *lexical semantic* relation exists. Further detail on the types of lexical cohesion is given below (section 2).

Lexical semantic relations are the building blocks of lexical cohesion, and so a clear understanding of their nature and behaviour is crucial, especially to computational applications of the theory, in information retrieval and automatic text analysis and understanding. *Classical relations*, the set of relations consisting of taxonomy (*bird/robin*), hyponymy (*tool/hammer*), meronymy (*hand/finger*), antonymy (*go/come*), and synonymy (*car/automobile*) are widely studied and applied. The focus in this paper, however, is on *non-classical* lexical semantic relations. These are the neglected relations that are not easily characterized as a sharing of the same individual defining properties between words. For example, *sail/wind* and *garden/digging* are instances of non-classical relations, whereas *car/auto* and *bird/robin* are not, since *car* and *auto* are related by sharing (nearly) all of the same properties, and a *robin* shares all of the properties of a *bird* and then has some unique ones of its own.[2] Clearly the same type of analysis cannot be made for either non-classical pair given above; however, both word pairs are obviously semantically related.

In lexical cohesion research in linguistics, and in implementations of it in computational linguistics (CL), non-classical relations are largely ignored. Most of the research on lexical semantic relations in linguistics, CL, and psychology has also ignored non-classical relations. A notable exception to this trend has occurred in library and information science (LIS). Most word pairs classed as RTs (related terms) in LIS thesauri are related non-classically, but unfortunately they are listed as an undifferentiated group. Further detail on non-classical relations is given below (section 3).

The research on types of lexical semantic relations in all disciplines has been done out of the context of text and then often assumed to be relevant within it, and no research has been done with human subjects identifying lexical cohesion in text. Therefore, we will conduct experiments with human readers of text (detailed below in section 4) to investigate these three major issues: the extent to which non-classical relations are involved in lexical cohesion, how relations are interpreted in the context of text, and the need for analysis of human readers' identification and interpretation of lexical cohesion.

A better understanding of reader-oriented relations[3] (i.e., the types of relations identified by readers of text) could potentially lead to improvements to relation types used in information

retrieval thesauri, and in lexical resources such as WordNet (Fellbaum, 1998) which is widely used (and free) in natural language processing applications. Lexical cohesion analysis has been used in automatic text analysis and understanding applications such as determining the structure of text (Morris & Hirst, 1991) and text summarization (Barzilay & Elhadad, 1999). Reader-oriented approaches will hopefully result in providing users with more natural and easy-to-use interfaces to text retrieval systems, thereby improving their information literacy skills.

2. LEXICAL COHESION

Approaches to lexical cohesion[4] in linguistic analysis are all based on Halliday & Hasan's fundamental idea of "continuity of lexical meaning". Two main types have emerged, distinguished by whether the groups of words that are joined together are further structured or not:

1. Unstructured lexical grouping
   - Unrestricted lexical continuity
   - Identity of reference
   - Similarity
   - Repetition
2. Structured lexical grouping
   - Cohesive harmony (Hasan, 1984)
   - Patterns of lexical affinities

In order to illustrate these types of cohesion, the following short story will be used. It was taken from a group of 80 such stories written by six- or seven-year-old children, analyzed by Hasan (1984, p. 189):

1. *there was once a little girl and a little boy and a dog*
2. *and the sailor was their daddy*
3. *and the little doggy was white*
4. *and they liked the little doggy*
5. *and they stroke it*
6. *and they fed it*
7. *and he run away*
8. *and then daddy had to go on a ship*
9. *and the children missed 'em*
10. *and they began to cry*

Unstructured lexical grouping refers simply to a group or list of related words, often referred to as a lexical chain. *Unrestricted lexical continuity* was the type of lexical cohesion originally defined by Halliday & Hasan (1976). Examples (with sentence numbers given) include these:

- 1 *little* 1 *little* 3 *little* 4 *little*
- 1 *girl* 1 *boy* 9 *children*
- 1 *dog* 3 *doggy* 4 *doggy*
- 1 *was* 2 *was* 3 *was*
- 2 *sailor* 8 *ship*

- 2 *daddy* 8 *daddy*
- 7 *run away* 8 *go*
- 9 *missed* 10 *cry*

Hasan (1984) separated lexical continuity chains into two different kinds: those involving identity of reference, and those involving similarity with no identity of reference. Her identity-of-reference chains consist of the following three examples:

- 1 *girl* 1 *boy* 2 *their* 4 *they* 5 *they* 6 *they* 9 *children* 10 *they*
- 2 *sailor* 2 *daddy* 4 *they* 5 *they* 6 *they* 8 *daddy* 9 *'em*
- 1 *dog* 3 *doggy* 4 *doggy* 5 *it* 6 *it* 7 *he* 9 *'em*

Furthermore, the *children* and *daddy* chains are joined through the pronoun *they* in sentences 4, 5, and 6, and the *daddy* and *doggy* chains are joined by the pronoun *'em* in sentence 9. Similarity chains from the example include these:

- 1 *little* 1 *little* 3 *little* 4 *little*
- 1 *was* 2 *was* 3 *was*
- 7 *run away* 8 *go*

Repetition chains have been studied by others (Hearst, 1997; Hoey, 1991) outside of the context of Halliday & Hasan's (1976) lexical cohesion, but nonetheless represent an analysis of continuity of lexical meaning.

All four of these chain-forming methods create groups of words exhibiting continuity of lexical meaning[5]. Hasan (1984), who introduced the idea of distinguishing identity and similarity chains, found the difference to be significant with respect to her goal of quantitatively correlating lexical cohesion with the human-ranked coherence of the stories she analyzed. When she combined the number of words involved in identity chains, similarity chains, and cohesive harmony (discussed below) she was successful in finding the correlation. The original lexical chains of Halliday & Hasan (1976) did not include identity of reference because it was classed as reference cohesion, which is not lexical cohesion but rather a form of grammatical cohesion. However, another of Hasan's goals was to create a cohesive environment in which "lexical and grammatical cohesion operates harmoniously" (1984, p. 203). A result of this approach is that the identity-of-reference chains could subsume chains which otherwise would have been lexically separate, causing a loss of any potential meaning contributed by lexical chains alone. For example, there are smaller separate lexical (only) chains for *children*, the *dog*, and the *daddy*, which become part of larger identity-of-reference chains. This can work the other way, as pointed out by Martin (1992, p. 427), where a longer similarity chain would not be formed if shorter pieces of it formed separate identity-of-reference chains. Hasan (1984) seems to give preference to identity of reference chains.[6]

Similarity chains are formed only by classical relations. In this way they differ substantially from the unrestricted lexical chains. All of the relations that Halliday & Hasan (1976, p. 285) originally referred to as "not easy to classify in systematic semantic terms" are excluded. In the story above, the chain consisting of *sailor* and *ship* is not a similarity chain, nor is the chain consisting of *missed* and *cry*. In the lexical chains created from five general-interest magazine articles by Morris & Hirst (1991), *Roget's Thesaurus* (1977) was used to include these hard-to-classify relations we refer to as non-classical. However, like the RTs listed in LIS thesauri, the

relations in *Roget's Thesaurus* are unclassified, so that whether a word pair is related but not how, can often be determined. Note that Hasan considers *stroke* and *fed* to form a similarity chain, but we do not. This issue is further discussed below.

The second main type of lexical cohesion, *structured lexical grouping*, involves a group of related words that has some further internal structure. In cohesive harmony (Hasan, 1984), identity-of-reference chains and/or similarity chains are linked together by grammatical intra-sentence relations similar to the case relations of Fillmore (1968)[7], such as agent/verb (*he* (the dog)/*run away,* from sentence 7 above) or verb/object (*stroke/it* (the dog), from sentence 5 above). The actual rule for group formation is that chains can be joined together if (at least) two instances of the same case relation exist between them. Hasan explains that "The source of unity ... resides in the fact that similar 'things' are said about similar/same 'entities', 'events', etc." (1984, p. 212). Texts with more words participating in cohesive harmony, and fewer chains left isolated, were consistently judged as more coherent. An example of two instances of a verb/object case relation, *stroke/dog* and *fed/dog*, joining together a similarity chain (*stroke, fed*) and an identity-of-reference chain (*it* (dog), *it*), occurs in sentences five and six in the above story. An example of two instances of an agent/verb case relation, *dog/run away* and *sailor/go*, joining together an identity of reference chain (*he* (dog), *daddy*) and a similarity chain (*run away, go*) occurs in sentences 7 and 8. Note that *dog* and *daddy* are related in the identity chain through the pronoun *'em*, used in sentence 9.

Cruse (1986) discusses a concept of *patterns of lexical affinities*, totally out of the context of Halliday & Hasan's (1976) lexical cohesion. However, the concept can be related to Hasan's (1984) cohesive harmony, but is more general in its application. He describes two types of affinities: *syntagmatic* and *paradigmatic*. A "syntagmatic affinity is established by a capacity for normal association in an utterance" (1986, p. 16). For example *dog* and *barked* are related in *The dog barked* (1986, p. 16), and *stroke* and *it* (dog) are related in sentence 5 of the story above. On paradigmatic affinity, he states "paradigmatically, a semantic affinity between two grammatically identical words is the greater the more congruent their patterns of syntagmatic normality" (1986, p. 16). His example is "*Arthur fed the dog/cat/?lamp-post*" (1986, p. 16). An example from the above story is "*They stroked/fed it*", which is also one of Hasan's examples of cohesive harmony. Cruse's paradigmatic affinities are more general than cohesive harmony, in that "patterns of syntagmatic normalities" are potentially more general than identical case relations, "grammatically identical words" are more general than identity of reference or similarity, and also because his intent seems to be that these affinities are acquired or established through general reading (i.e., through relations between texts), and he does not apply them to the context of a specific text. However, the story example given above applies, and is revisited below in section 3, in the discussion on whether relations are text-specific.

Consider a text that includes the following three (constructed) sentences:

Arthur fed the dog.
That stupid boy did not feed his cat.
He stroked the gerbil.

These sentences all fit the syntagmatic pattern of human agents (*Arthur*, *boy,* and *He*) doing something to objects that are pets (*dog*, *cat*, and *gerbil*), in the case sense of agent and object. This analysis applies a combination of similarity grouping (potentially more general than

Hasan's 1984 groupings) and an inter-sentence pattern recognition based on identical intra-sentence case relation structure that is the same as the principle behind cohesive harmony. The example illustrates both the potential for learning relations ("semantic affinities") from texts as Cruse (seemingly) intended, as well as potential application within a specific text, since although the example is made up, one could imagine all three sentences existing within the same text.

Cohesive harmony (Hasan, 1984) and the concept of patterns of lexical affinity (Cruse, 1986) make the important contribution of linking lexical inter-sentence cohesion with grammatical intra-sentence cohesion. The result is that more tightly knit (and structured) groups of related words can be formed within text. Hasan (1984, p. 218) concludes that "cohesive harmony is the lexico-grammatical reflex of the semantic fact of coherence". Both of these concepts also bring together text-specific cohesion (identity of reference and intra-sentence case relations) with general (out of the context of a text) cohesive devices such as Halliday & Hasan's (1976) original concept of lexical cohesion, Hasan's similarity chains, and Cruse's paradigmatic semantic affinities that are assumed to exist generally within the system of language. This could be described as harmonizing lexico-grammatical continuity within a specific text and lexico-grammatical continuity with other texts, which is similar to, but more general than Hasan's concept of harmonizing lexical and grammatical cohesion within the same text. This aspect of harmony is more general than Hasan's, because like Cruse's paradigmatic affinities, it is expressly presented as interacting with other texts. Perhaps the same applies to cohesive harmony, but it is not presented as doing so, other than the fact that the word pairs generated by relations in similarity chains are not text-specific, and the existence of a general assumption that language output is affected by language input. These concepts have not been applied computationally, and it remains to be seen if they will be reflected in the cohesion identified by readers in the forthcoming experiments (section 4 below). Ultimately, an expanded concept of cohesive harmony that would include lexical, grammatical, and semantic linguistic features, linking together words, sentences, and inter-sentence units such as lexical chains (and/or other groupings), both within a text and between texts, would appear to be a useful goal.

One consequence of cohesive harmony and patterns of lexical affinity is that they depend on noun/verb separation. In fact, for Hasan (1984), this is a more general result of using only classical lexical semantic relations. Being based on similarity, as they are, means that only identical word classes can be related. Cruse (1986) stipulates that affinities will occur only between like grammatical classes. The lexical chains of Morris & Hirst (1991) had no such restriction, and frequently nouns, adjectives, adverbs, and verbs were joined together in one chain. Again, it will be interesting to analyze readers' lexical chains in this regard.

Finally, it must be noted that Hasan's (1984) analysis was based on the relatively simple narrative structure of children's stories. Martin (1992) applied the theory successfully to a teenager's story, and Hasan (1984, p. 218) states that "this kind of analysis has been applied to data from other genres, and the hypotheses regarding the relation between cohesive harmony and coherence have not been challenged".

3. NON-CLASSICAL LEXICAL SEMANTIC RELATIONS

Consider the relations between the following words, a portion of a lexical chain from a general interest magazine article (Morris & Hirst, 1991, p. 43): *suburbs, driving, car's, lights, drive, urban,* and *traffic*. These are mostly non-classical relations, and it is to this topic that we now turn. From an LIS perspective, the word pairs can potentially be related via the non-hierarchical RT or "associative" relation, as shown in the following (constructed) example of thesaural relations for *car*:

- Classical: BT *vehicle* NT *sedan* (hyponymy/taxonomy) UF *automobile* (synonymy)
- Non-classical: RT *drive, traffic*

However, as stated earlier, the specific non-classical relations cannot be determined. This is also the case with *Roget's Thesaurus* (1977), which was used to form the above chain fragment. Although this thesaurus is hierarchically classified, it makes frequent use within its basic categories, of unclassified pointers to other widely dispersed basic categories. In this respect the structure of LIS thesauri and *Roget's Thesaurus* are similar. They are both hierarchically organized, *Roget's* by Roget's own principles of domain/topic division, and LIS thesauri by the BT/NT (broad term/narrow term) structure, but they also both have a non-hierarchical, non-classified "structure" (or at least mechanism) for representing non-classical relations.

Lakoff (1987) gives the name "classical" to categories that are related because their members all share the same common properties.[8] This comes from the classical views of Aristotle, that these common properties (plus a differentia) will be necessary and sufficient for category definition. Following this terminology, we will refer to relations that depend on the common properties of classical categories as "classical" relations. Hence the term "non-classical"[9] for relations that don't depend on the common properties required of classical relations.[10] One of the major points made by Lakoff is the importance of non-classical *categories*, providing support for the importance of non-classical *relations*. The classical category structure has been a limiting medium for relations. It follows that since relations create categories (or vice versa), if categories are severely restricted in nature, so too will be the relations. Classical relations are restricted in that they only deal with similarity relations between words in the same grammatical class. The following are the major (not necessarily mutually exclusive) types of non-classical relations:

- Relations between members of Lakoff's non-classical categories: *ball, field,* and *umpire*, that are part of the structured activity of *cricket* (or *baseball*)
- Case relations:
  - General: *dog/bark* (Chaffin & Herrmann, 1984)
  - Sentence-specific (Fillmore, 1968): *stroke/it* from the above story
- LIS RTs (Milstead, 2001)

The relations between members of non-classical categories are interesting in that in many cases they are unnamable except with reference to the category name (*ball/field* or *ball/umpire*, without using the word *cricket*). For word pairs consisting of a member and the category name, the relation has often been covered, either as a general case relation (*ball/cricket* as instrument/activity) or as an RT (*field/cricket* as the activity/location relation of Neelameghan (2001), or the locative general case relation).

Case relations come in two varieties: general and specific (to a sentence). The general inter-sentence and inter-text case relations (Chaffin & Herrmann, 1984) are given also by several of the LIS researchers who have provided lists of RT types (Neelameghan, 2001; Milstead, 2001). Cruse deals almost exclusively with classical relations, but does mention two general case-like relations he calls "zero-derived paronymy" (1986, p. 132). The instrumental case (*dig/spade* or *sweep/broom*) and the objective case (*drive/vehicle* or *ride/bicycle*) are given as examples. He observes that in the instrumental case, the definition of the noun will most likely contain the verb, and in the objective case, the definition of the verb will most likely contain the noun, which could be relevant computationally. To Cruse, these relations are not "real" relations, but rather "quasi" relations, since the word classes involved differ. Neither Hasan (1984) or Martin (1992) consider them as relations contributing to lexical cohesion.

The case relations as defined by Fillmore (1968), and the case-like relations used by both Hasan (1984) and Martin (1992) in lexical cohesion research, are intra-sentence grammatical relations that always apply to the specific text and sentence they are situated in. Sometimes these relations can be both text-specific and general at the same time (*dog/barked* in *The dog barked*).

LIS can lay claim to the most extensive amount of research on non-classical relations. This is likely a pragmatic reflection of the fact that it is a field with a large user base that demanded this type of access to reference materials long before computer access was feasible. An example of this user demand happened during the development of the *Art and Architecture Thesaurus* (AAT), where RTs were not included in the initial design, but rather added afterwards due to user demand. Standards for their use have been developed (ISO, 1986); however, the Library of Congress has been encouraging a minimization of their use since 1985 (El-Hoshy, 2001). Since RTs are all grouped together in an unclassified manner, the result has been inconsistencies and subjective judgements about what word pairs get included, and as such, is an implementation issue rather than a usefulness issue.

Neelameghan (2001) has produced the most extensive list of non-classical relations, which has changed little since 1976 (Neelameghan & Ravichandra). Apart from relations between members of non-classical categories (see above), his list includes most of the text-general relations (recognizable out of the context of a text) mentioned by other researchers. Obviously any text-specific relations such as sentence-specific case cannot be included, since word pairs are considered out of text. A couple of exceptions have been noted. Evens et al. (1983) have a *provenience* relation (*water/well*), and Cruse (1986) has a *proportional series* relation made up of recurring *endonymy* (*university/lecturer/student*, *prison/warden/convict*, *hospital/doctor/patient*). Endonymy is a relation that "involves the incorporation of the meaning of one lexical item in the meaning of another", such as *pig* in *bacon/ham/sty* (1986, p. 123–125).

Now that the three main types of non-classical relations have been discussed, we turn to some general issues. An important question regarding non-classical relations is why they have been so neglected. Part of the answer is historic. Aristotle defined classical categories, which resulted in classical relations, which created the research focus on them. This research focus has been noted by Cruse, who focuses on the classical relations since "A relatively small number of these have come to occupy focal positions in discussions of lexical semantics (such relations as antonymy, hyponymy, and synonymy)" (1986, p. 85–86). Hasan (1984, p. 213) refers to her usage of

"readymade" categories from linguistics. The classical relations of taxonomy and hyponymy (and also meronymy, but it is less straightforward) are hierarchical, allowing elegant formal organizing structures with inheritance properties to be created. Non-classical relations seem messy and unorganized by comparison.

Classical relations are assumed to exist within the system of language (Hasan, 1984). Perhaps because we are culturally trained to think in terms of their primacy, it is hard to think of *beer/baseball*, *ball/field*, or *bread/butter* as being related within the system of language, even though some psychological lexical-priming experiments (Hodgson, 1991) indicate that the involved relations are as strongly perceived by humans as the classical relations are. Rather, these non-classical relations appear to be somehow located within the system of the world, not the system of language, even though there does not seem to be a fundamental linguistic reason for this. The difference between them and the classical relations could just be that because the classical ones have been named and used for so long in academic analysis, they are seen to be part of the language system, as distinct from the world system. The general case relations seem to fall somewhere in-between on this issue, perhaps because they seem (or in some way are) grammatical like Fillmore's (1968) case concept, and the purely lexical non-classical relations are not considered as seriously. There is an analogy to this in psychology, where in lexical-priming experiments, "associative relations", operationalized as word associations, were given a much lower status than "semantic" relations, which are classical.

Hasan (1984, p. 195) remarked that if relations other than the classical ones were used, inter-subjectivity would become a problem. However, classical relations can also be inter-subjective. While she considers *stroke/fed* and *liked/missed/cry* to be related classically (presumably by either co-hyponymy, or co-meronymy), it is unlikely that agreement is universal. One of the goals of the our experiments is to investigate inter-subjectivity.

Another related issue is whether relations are text-specific or text-general. This is an important issue, since text-general relations should be "findable" or "put-able" in a resource, whereas text-specific relations will require further potentially complex interactions with the rest of the text for their identification. Hasan (1984, p. 201) claims that classical relations are "dissociated from a real context of utterance". However, consider again the supposed classical relation between *stroke* and *fed* from the story above. The relation must either be co-hyponymy or co-meronymy. For the purposes of this discussion we will assume it is co-hyponymy, where the parent category might be "things done to dogs" or "what people often do to pets". Hasan's implementation of cohesive harmony depends on the classical relation pre-existing out of the context of the story, but the relation seems specific to the text, due to repeated mention of *dog*, and the fact that *dog* stands in an objective case relation to both verbs. Therefore, rather than the classical relation between *stroke* and *fed* pre-existing, it seems that part of the mechanism of cohesive harmony itself causes the two words to become related.

This example is reminiscent of Barsalou's (1989) concept of creating ad hoc categories, his term for categories that are "made up on the fly for some immediate purpose", which would presumably require a similar type of interaction with a specific text, instead of the assumption that all categories pre-exist (Lakoff, 1987, p. 45).[11] Two examples of these categories are "things to take on a camping trip" and "what to do for entertainment on a weekend" (1987, p.

45). Barsalou's ad hoc categories seem to fall into (at least) two types: (1) different activities or actions pertaining to the same or similar objects; (2) different objects pertaining to the same or similar activities or actions. The above example of *stroke/fed* could be an example of a category being created because of different actions relating to the same thing, via the objective case relation. An example of the second type is "things to take on a camping trip". Since categories created this way are not classical, as they seem to be ways of joining "different" objects, actions, or activities, the relations between their members are not classical either. As in cohesive harmony, and with patterns of lexical affinities, the mix of classical categories and relations with non-classical categories and relations seems to be a rich source of lexico-grammatical cohesion.

Barrière & Popowich (2000) have carried out some related research, using relations between specific verbs and specific case role fillers (e.g., "things that *carry people*"), that create non-classical categories. They propose adding these new categories to a classical hierarchy automatically, using a children's dictionary that often employs case relations to define words, as long as the number of instances that match each pattern is greater than a pre-specified threshold. Therefore, the potentially enormous numbers of categories and relations so produced are assumed to pre-exist, at least implementationally. The categories are described as being non-lexical, in that (presumably) no word exists that is suitable as the category name. Their categories have a common hierarchical property in that the more specific they become (i.e., the larger the number of specific relations involved in their definition), the further down in the hierarchy they go. For example, "things that *carry*" is the parent of "things that *carry water*" and "things that *carry people*". This method is potentially a source of non-classical relations, but whether they are useful or used by readers remains to be determined. Perhaps in some cases, the categories are non-lexical because they are rarely useful in practice, or would be better implemented in a text-specific way such as with ad hoc categories à la Barsalou.

Certainly many of the classical relations require minimal context for their identification. The fact of shared and identical properties means that given one word of a classically related pair, all related words are implied or known, even without the actual presence of a related word. For example, given *bird*, all creatures below it in the taxonomy are known to be *bird*-like; given *automobile*, all synonyms such as *car* are implied. Therefore, for many classical relations, not only is surrounding text, as context, not needed, but the other related word of the pair is not needed either. They are truly stand alone in terms of the context they require, which is probably why they are so often implemented. This is not true of all non-classical relations, although some words (*broom* in *broom/sweep*) seem strongly suggestive of related words. In this case, *sweep* is part of the definition of *broom*. For relations between members of a non-classical category (*plow/cow*), both words are needed to form the relation, unless a category such as "things that may occur on a farm" is assumed to exist before the first word is seen in text.

A property of classical relations is *semantic distance*. Semantic distance is often measured as the number of links between words in a hyponymy hierarchy, or variants thereof (Budanitsky & Hirst, 2001), and is a measure of conceptual closeness due to identical shared properties. For example, *robin* and *bird*, being fewer links apart, would be more "closely" related than *robin* and *animal*, where there are more intervening links. On the other hand, non-classical lexical semantic relations clearly will not exhibit this property of semantic distance the same way, since the relations themselves are not based on common shared properties. The non-classical relations

are lateral (Neelameghan, 2001) or non-hierarchical. Once hierarchical classical relations are mixed with non-hierarchical non-classical relations, the current measures of classical semantic distance will require modification.

One of the goals of this research is to investigate how far domain-neutral relations can go in relating non-classically related words in text, without resorting to hundreds (Cassidy, 2000) or thousands (Lenat, 1995) of relations, or perhaps even more in the case of Barrière & Popowich (2000). In other words, is there a smallish set of field- (domain-) neutral relations that will provide (good) coverage for all (or most) fields? In this context, the issue of field-specific versus field-neutral word pairs and relations will be discussed. Encouragingly, LIS has tackled an extensive number of specific domains with just such a smallish set of field-neutral non-classical relations. However, due to the reportedly subjective implementation of these relations, this may not in fact be true in practice. Martin (1992) suggested that taxonomic relations can deal with specific domains, but despite the unlikelihood of this, his view suggests a pragmatic computer implementation of determining field/domain as quickly as possible, and then mining the relevant LIS thesauri (or other available resources). WordNet's approach uses field-neutral relations for a general domain, but mostly for classical relations. Databases use field-specific relations for specific domains. The experiments to be described below will, it is hoped, provide some insight.

4. PLANNED EXPERIMENTS

4.1 Problem Statement

The investigation of lexical cohesion and lexical semantic relations in the preceding three sections has led to the identification of three issues. First, non-classical relations are ignored in recent lexical cohesion research and simply mentioned as a difficult group in the original Halliday & Hasan (1976) research. They have been ignored in most recent research in linguistics, CL, and psychology (for a call to broaden the use of relation types, see McRae & Boisvert, 1998), but studied as a group in LIS. Different LIS researchers have created lists of them, but since they are grouped together indiscriminately in thesauri, it is not clear whether the lists have been adhered to. Second, lexical semantic relations have not been studied in the context of text. Usually, word pairs are given as examples, out of a specific textual context. In lexical cohesion research (obviously in the context of text) only classical relations are used, as well as specific (Fillmore, 1968) case relations. Third, lexical cohesion has not been studied with a group of human participants (who are not linguists studying lexical semantic relations) as readers of text, to see what types of lexical cohesion they use, and also what lexical semantic relations they use. Most research is done using researcher-defined types. Spiteri (2002) proposed studying user-defined relations, but by using word-association methodology, out of the context of text. No research has been done on what the lexical chains mean or signify as units with respect to the text in which they are situated.

4.2 Primary Objectives

The primary objective of the experiments is to analyze the lexical semantic relations in the lexical chains identified in text by subjects. The following questions will be addressed:

- How similar are the lexical chains among the participants?
- What kinds of reader-defined relations are used in the lexical chains, including whether they are non-classical or classical, and grammatical or lexical?
- Do the participants agree on how the identified word pairs are related?
- Is there an overall set of similarly named or described relations identified in the texts?
- What kinds of word pairs are used, including whether they are text-specific or text-general, and what word classes are used?
- Can lexical chains, which are examples of text-specific categories, be similarly named or described by participants?

It is anticipated that participants will identify similar lexical chains in the text, and that both classical and non-classical lexical semantic relations will be used. It is also anticipated that participants will be able to provide similar names or descriptions for the lexical chains in the texts, assuming that the chains themselves are statistically similar among participants. As well, we anticipate that participants will be able to provide similar names or descriptions for the same relations (i.e., the same word pairs) identified in the texts, assuming that statistically similar word pairs are used in forming the chains in the first place, giving identical word pairs for analysis.

4.3 Methodology

An experiment will be carried out with approximately 30 participants (from within the university student population) marking the lexical cohesion (i.e. the lexical chains) in the text. Participants will also indicate for each related word pair, how the words are related by giving a name or short description. They will indicate what each chain means by giving a name or short description. These lexical chains and relation descriptions will then be analyzed to answer the questions above given as primary objectives.

The texts used will consist of the first page of text of three general-interest articles from current *Reader's Digest* magazines. Participants will mark the chains (e.g., by underlining words) while reading the text. They will be instructed not to hurry, but to read naturally for comprehension, not studying the text for lexical chains. When they finish marking all of the chains, they will divide the words into separate chains. They will be asked to indicate why they put each word in the chain by indicating which word pairs are related. The instructions given for how to create the chains must not include specific examples of related word pairs; rather, an example with nonsense text showing chain markings and the subsequent separate chains in columns will be used.

Statistical analysis will be used for determining similarity among the participants' chains and similarity among the names or descriptions of relations and chains. Conceptual content analysis will first be used to identify names or descriptions that can be considered identical. The relations will be classified as to what word classes are used in the word pairs, and also as to whether they are text-specific or text-general. If necessary, participants other than those who originally identified the relations could be used to judge this specificity.

5. DISCUSSION

To automate the detection of lexical semantic relations between words in text, an inventory of the types of relations *that are used by readers* is needed. Our experiments will provide data on the usage of classical and non-classical relations in text, and on how they contribute to lexical cohesion in text.

There is currently no information on whether participants identify similar lexical chains in text, nor on whether they identify similar relations. These experiments will provide data that can be used to get a sense of the strengths and limitations of analyzing the lexical cohesion in text. They will provide information that could potentially be used in building lexical resources that contain reader-oriented relations, or indicate where text-specific analysis is necessary. Such resources could be used in information retrieval to augment user queries and find the most relevant textual documents. The use of such lexical resources could also be a valuable aid to the automatic detection of lexical cohesion in text, which contributes to aspects of its meaning.

We now turn to discussing potential problem areas and/or limitations of this research. Reading comprehension is an extremely complex process, and lexical cohesion is but one aspect of this process. When participants are asked to identify and mark lexical chains, does this interfere in a significant way, with the natural process of comprehending the cohesion in the text? Lexical cohesion creates lexical chains that can be considered to be semantic units (Hasan, 1984). In these experiments, participants are asked about relations between individual word pairs. This does not directly address the possibility of words going into the chain because of a relation to this unity, rather than to individual words. Participants could be asked for the relation information at the time of adding a word to a chain, but it would have to be determined if this would affect which words go into the chains, or any other significant aspect of chain formation. We also acknowledge the following practical limitations: only three short texts are used, only the general-interest, light-reading genre is used, and divergent kinds of participants are not used.

Finally, we indicate some potential future research areas, several of which suggest ways of overcoming these limitations. Rather obviously, one could explore the ramifications of using different types of participants, different genres, and longer samples of text. Further research could be done on lexical chains as units, to determine whether words enter chains because of relations to the chain as a unit, or because of relations to specific words within the chain.

Insight may be gained from qualitative approaches, such as working with participants closely as they create chains, eliciting from them how they are doing it, and asking them questions about the process. If participants have problems naming relations, then the approach of asking them to choose relations from a pre-defined set could be used. The question of what pre-defined set to use could potentially be aided by qualitative analysis of participants' focused discussion on related word pairs, combined in some way with the relations that have already been defined by researchers. It is certainly anticipated that it will be easier for participants to identify which word pairs are related than to explain how they are related.

Further research could be done into how Barsalou's (1989) ad hoc categories are created in the context of lexical cohesion. Automating cohesive harmony analysis, and the inclusion of

reference cohesion with lexical cohesion, would highlight benefits over current implementations of lexical cohesion that work solely with similarity chains. Similarly, investigating patterns of lexical affinity could have computational benefits, in both text-specific and text-general analysis.

In summary, this paper advocates a reader-oriented approach to the analysis of the lexical semantic relations that contribute to lexical cohesion in text. Lexical cohesion, in turn, contributes to the semantic understanding of the text. We anticipate that this approach will contribute to information literacy by helping to provide easier-to-use interfaces to systems that provide intelligent semantic access to text-based information.

ACKNOWLEDGEMENT

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

ENDNOTES

[1] By *structural*, Halliday & Hasan mean in the prescriptive sense of something like sentence grammar. They acknowledge work on discourse structure, but view it as a loose structure of a different kind.

[2] Antonymy and meronymy are classical, but slightly more difficult to characterize.

[3] Reader-oriented relations are analogous to the user-oriented relations proposed by Spiteri (2002).

[4] Not all of the approaches are referred to as lexical cohesion. This is Halliday & Hasan's (1976) term, but we use it to include all analyses that are based on continuity of lexical meaning.

[5] It is beyond the scope of this paper to discuss differences of, and the significance of repetition.

[6] This may actually be solely for the purpose of her quantitative goal of finding a correlation between cohesion and coherence, since she alludes to reasons for keeping lexical cohesion separate (Hasan, 1984, p. 197–199).

[7] Hasan (1984) identifies six case-like relations that she refers to as functional relations. Martin (1992) expands on this group, and notes the similarity to case relations (1992, p. 294). He calls the result *nuclear relations*.

[8] We acknowledge that Lakoff (1987) is referring to categories in the general conceptual sense, and we are referring to categories represented by words. This does not affect our analysis.

[9] The LIS term *associative* is not used because of its strong association with word-association in psychology and elsewhere.

[10] For the words in classical relations, the properties that are used to define the words or concepts as individuals are then shared (or explicitly not shared, as in antonymy). This is a very individualistic view of both relations and words, and the classical relations actually seem to be the most "un-relation-like" of the relations, dealing with definitional (rather than relational) aspects of words.

[11] For things that become a significant-enough part of one's personal or cultural experience, the category may become fixed, leading to such expressions as "*Beer* and *baseball* are synonymous for me", "*Beer* and *baseball* certainly go together", "*baseball* and *apple pie*", and "I cannot imagine *hockey* without *fighting*".

REFERENCES

Barrière, C., & Popowich, F. (2000). Expanding the type hierarchy with nonlexical concepts. In H. Hamilton (Ed.), *Canadian AI 2000* (pp. 53–68). Berlin: Springer-Verlag.

Barsalou, L. (1989). Intra-concept similarity and its implications for inter-concept similarity. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 76–121). Cambridge: Cambridge University Press.

Barzilay, R., & Elhadad, M. (1999). Using lexical chains for text summarization. In I. Mani & M. Maybury (Eds.), *Advances in text summarization* (pp. 111–121). Cambridge, Mass.: MIT Press.

Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *WordNet and other lexical resources: Applications, extensions, and customizations, NAACL 2001 Workshop* (pp. 29–34). Association for Computational Linguistics.

Cassidy, P. (2000). An investigation of the semantic relations in the Roget's Thesaurus: Preliminary results. In A. Gelbukh (Ed.). *CICLing-2000: Conference on Intelligent Text Processing and Computational Linguistics, February 13–19, Mexico City*, 181–204.

Chaffin, R., & Herrmann, D. (1984). The similarity and diversity of semantic relations. *Memory and Cognition,* 12(2), 134–141.

Cruse, D. (1986). *Lexical semantic relations.* Cambridge: Cambridge University Press.

El-Hoshy, S. (2001). Relationships in Library of Congress Subject Headings. In C. Bean, & R. Green (Eds.), *Relationships in the organization of knowledge* (pp. 135–152). Norwell, Mass: Kluwer Academic Publishers.

Evens, M., Markowitz, J., Smith, R., & Werner, O. (Eds.). (1983). *Lexical semantic relations: A comparative survey.* Edmonton, Alberta: Linguistic Research Inc.

Fellbaum, C. (1998). *WordNet: An electronic lexical database.* Cambridge, Mass.: MIT Press.

Fillmore, C. (1968). The Case for Case. In E. Bach & R. Harms (Eds.), *Universals in linguistic theory* (pp. 1–88). New York: Holt, Rinehart & Winston.

Halliday, M.A.K., & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Hasan, R. (1984). Coherence and Cohesive Harmony. In J. Flood (Ed.), *Understanding reading comprehension: Cognition, language and the structure of prose* (pp. 181–219). Newark, Delaware: International Reading Association.

Hearst, M. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics,* 23(1), 33–64.

Hodgson, J. (1991). Informational constraints on pre-lexical priming. *Language and Cognitive Processes,* 6(3), 169–205.

Hoey, M. (1991). *Patterns of lexis in text.* Oxford: Oxford University Press.

ISO. (1986). Guidelines for the establishment and development of monolingual thesauri. [Geneva:]. ISO. (ISO2788-1986(E)).

Lakoff, G. (1987). *Women, fire and dangerous things.* Chicago: University of Chicago Press.

Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM,* 38(11), 33–38.

Martin, J. (1992). *English text: System and structure.* Amsterdam: John Benjamins.

McRae, K., & Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 24(3), 558–572.

Milstead, J.L. (2001). Standards for relationships between subject indexing terms. In C.A. Bean and R. Green (Eds.). *Relationships in the organization of knowledge* (pp. 53–66). Kluwer Academic Publishers

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics,* 17(1), 21–48.

Neelameghan, A. (2001). Lateral relationships in multicultural, multilingual databases in the spiritual and religious domains: The OM Information Service. In C. Bean & R. Green (Eds.), *Relationships in the organization of knowledge* (pp. 185–198). Norwell, Mass.: Kluwer Academic Publishers.

Neelameghan, A., & Ravichandra, R. (1976). Non-hierarchical associative relationships: Their types and computer generation of RT links. *Library Science,* (13), 24–42.

Roget, P. (1977). *Roget's international thesaurus* (2nd ed.). Harper & Row Publishers Inc.

Spiteri, L. (2002). Word association testing and thesaurus construction: Defining inter-term relationships. In *Proceedings of the 30th Annual Conference of the Canadian Association for Information Science*, Toronto, Ontario, May 30–June 1, 24–33.