Lexical Cohesion, the Thesaurus, and the Structure of Text

Jane Morris

Technical Report CSRI-219
December 1988

Computer Systems Research Institute
University of Toronto
Toronto, Canada
M5S 1A1

Lexical Cohesion, the Thesaurus, and the Structure of Text

by

Jane Morris

Department of Computer Science
University of Toronto
Toronto, Ontario, Canada

September 1988

A Thesis submitted in conformity with the requirements
for the degree of Master of Science

Abstract

In text, lexical cohesion is the result of chains of related words that contribute to the continuity of lexical meaning. These lexical chains are a direct result of units of text being about the same thing. Finding text structure involves finding units of text that are about the same thing. Hence, computing the chains is useful since they will have a correspondence to the structure of the text. Determining the structure of text is an essential step in determining the deep meaning of the text. In this thesis, a thesaurus is used as the major knowledge base for computing lexical chains. Correspondences between lexical chains and structural elements are shown to exist. Since the lexical chains are computable, and exist in non-domain-specific text, they provide a valuable indicator of text structure.

Acknowledgements

I would like to thank Graeme Hirst for his tremendous level of support, guidance, and encouragement. Thanks to Robin Cohen, Jerry Hobbs, Eduard Hovy, and Ian Lancashire for valuable discussions of the ideas in this thesis. Thanks to Chrysanne Di Marco, Mark Ryan, and John Morris for proof-reading, and commenting on drafts of this thesis. A special thanks to John Morris for being truly supportive of this work.

The opportunity I was given to study computer science at the University of Toronto was financially assisted by the Government of Ontario and the Department of Computer Science of the University of Toronto.

For permission to reproduce copyright material in this report, I am grateful to Dr A.M. Clarfield, Dr Adrian Forsyth, Jay Teitel, *Equinox* magazine, *The New Yorker*, Reader's Digest Magazines Limited, and *The Washington Post*.

Contents

Chapter 1

Lexical Cohesion

This thesis is about identifying and tracing patterns of lexical cohesion called *lexical chains*. These patterns can then be used for word interpretation in context, including both narrowing to specific shades of meaning, and word sense disambiguation. Since these lexical patterns are a natural consequence of organized text, their determination serves as an indicator of the structure of the text containing them. This work provides one mechanism that must be integrated into the total machinery required for a computational text understanding system.

The thesis covers the following topics:

1. What lexical cohesion is and why it is important.

2. The thesaurus and lexical cohesion.

3. How to find lexical chains.

4. What to do with lexical chains.

1.1 What Lexical Cohesion Is

In this section, lexical cohesion will be defined by briefly defining the general concept of cohesion, and then narrowing to the specific type of cohesion called lexical cohesion.

1.1.1 Cohesion

A text or discourse is not just a set of sentences, each on some random topic. Rather, the sentences and phrases of any sensible text will tend to be about the same things—that is, the text will have a quality of unity. This is the property of *cohesion*—the sentences "stick together" to function as a whole. Halliday and Hasan [5] have identified five characteristics of texts that contribute to cohesion:

- Reference: text items that refer to some other item in the text e.g., pronouns.

- Substitution: replacement of one text item by another, e.g., replacing a *noun* by *one*.

- Ellipsis: replacement of a text item by nothing.

- Conjunction: text items that express meanings that presuppose other items in the text such as *therefore* or *as a result.*
- Lexical cohesion: semantic word relationships.

The following example of cohesion of the reference type characterizes what is meant by a cohesive relation. In reference, the required **dependency relation** between two elements in the text is identity of reference. Put another way, both elements refer to the same thing. Consider the following simple example:

(1-1) 1. *John* is a nice guy.
2. I like *him.*

The interpretation of *him* is dependent on the interpretation of *John*, and the dependency relation is identity of reference, the common referent being *John.* Furthermore, the text functions as a unit, and not as two disjoint sentences.

Cohesion is not a guarantee of unity in text but rather a device for creating it. As aptly stated by Halliday and Hasan [5], it is a way of getting text to "hang together as a whole".

1.1.2 Lexical Cohesion

Now that the general concept of cohesion is defined, the type known as lexical cohesion can be defined.

Lexical cohesion is obtained through chains of related words, providing "continuity of lexical meaning" [5]. The dependency relation required here is simply that there be a **recognizable relation** between the words. To illustrate this point five different types of word relations are given with examples:

1. Reiteration with identity of reference:

   (1-2) (a) Mary bit into a *peach.*
   (b) Unfortunately the *peach* wasn't ripe.

2. Reiteration without identity of reference:

   (1-3) (a) Mary ate some *peaches.*
   (b) She likes *peaches* very much.

3. Reiteration by means of superordinate:

   (1-4) (a) Mary ate a *peach.*
   (b) She likes *fruit.*

4. Systematic semantic relation (systematically classifiable):

   (1-5) (a) Mary likes *green* apples.
   (b) She does not like *red* ones.

5. Non-systematic semantic relation (not systematically classifiable):

(1-6) (a) Mary spent three hours in the *garden* yesterday.
(b) She was *digging* potatoes.

Each example of lexical cohesion above shows a different type of word meaning relationship.

1.1.3 Classes of lexical cohesion

Halliday and Hasan [5] have provided a useful and complete classification of lexical cohesion based on the type of dependency relationship that exists between words. This classification will now be described using the above examples.

Examples 1-2, 1-3, and 1-4 fall into the class of *reiteration*. Note that reiteration includes not only identity of reference or repetition of the same word, but also general nouns {*man, idiot*}, superordinates, subordinates, and synonyms.

Examples 1-5 and 1-6 fall into the class of *collocation*. Collocation is a term for semantic relationships between words that tend to co-occur. The semantic word relationship is recognizable because the words are located near one another. There is a *distance* relation between the words, and the words co-occur within a given *span*. It can be further divided into two categories of relationship: *systematic semantic*, and *non-systematic semantic*.

Systematic semantic relationships are (as the name implies) those that can be *classified* in a systematic semantic way. This type of lexical cohesion includes antonyms, members of an ordered set {*one, two, three*}, members of an unordered set {*white, black, red*}, and part-to-whole relationships {*eyes, mouth, face*}.

Example 1-6 is an illustration of collocation where the word relationship is a non-systematic semantic one. As a result, this type of relationship is the most problematic, especially from a knowledge representation point of view.

As stated above, collocation is a class of relationships between words that tend to occur in similar lexical environments. Words tend to occur in similar lexical environments because they tend to occur in similar situations or contexts. As a result, context-specific examples such as the set {*post office, service, stamps, pay, leave*} must be included in this class. This example is from Ventola [22], who analysed the patterns of lexical cohesion specific to the context of service encounters. Another example of this type is {*car, lights, turning*} which is taken from example (4-1) in chapter 4. These words are related in the situation of driving a car, but taken out of that situation, they are not related in a systematic way.

Also contained in the class of collocation are word associations. Examples from Postman and Keppel [16] are: {*priest, church*}, {*citizen, U.S.A*}, and {*whistle, stop*}. Again, the point can be made that the exact relationship between these words is hard to classify, but there does exist a recognizable relationship.

1.2 Lexical Chains

Often lexical cohesion occurs not simply as pairs of related words but as a succession of any number of related words spanning a topical unit of the text. These groups of related words will be called **lexical chains**.

Lexical chains are important, since they tend to delineate portions of text that have a strong unity of meaning. To illustrate the concept of lexical chains, consider this example (the first 16 sentences of example (4-1)):[1]

(1-7) 1. ¶I spent the first 19 years of my life in the suburbs, the initial 14 or so relatively contented, the last four or five wanting mainly to be elsewhere.

2. The final two I remember vividly: I passed them driving to and from the University of Toronto in a red 1962 Volkswagen 1500 afflicted with night blindness.

3. The car's lights never worked—every dusk turned into a kind of medieval race against darkness, a panicky, mournful rush north, away from everything I knew was exciting, toward everything I knew was deadly.

4. I remember looking through the windows at the commuters mired in traffic beside me and actively hating them for their passivity.

5. I actually punched holes in the white vinyl ceiling of the Volks and then, by way of penance, wrote beside them the names and phone numbers of the girls I would call when I had my own apartment in the city.

6. One thing I swore to myself: I would never live in the suburbs again.

7. ¶My aversion was as much a matter of environment as it was traffic—one particular piece of the suburban setting: the "cruel sun."

8. Growing up in the suburbs you can get used to a surprising number of things—the relentless "residentialness" of your surroundings, the weird certainty you have that everything will stay vaguely new-looking and immune to historic soul no matter how many years pass.

9. You don't notice the eerie silence that descends each weekday when every sound is drained out of your neighbourhood along with all the people who've gone to work.

10. I got used to pizza, and cars, and the fact that the cultural hub of my community was the collective TV set.

11. But once a week I would step outside as dusk was about to fall and be absolutely bowled over by the setting sun, slanting huge and cold across the untreed front lawns, reminding me not just how barren and sterile, but how undefended life could be.

12. As much as I hated the suburban drive to school, I wanted to get away from the cruel suburban sun.

13. ¶When I was married a few years later, my attitude hadn't changed.

14. My wife was a city girl herself, and although her reaction to the suburbs was less intense than mine, we lived in a series of apartments safely straddling Bloor Street.

15. But four years ago, we had a second child, and simultaneously the school my wife taught at moved to Bathurst Street north of Finch Avenue.

16. She was now driving 45 minutes north to work every morning, along a route that was perversely identical to the one I'd driven in college.

[1] ©1987 Jay Teitel. Reprinted by kind permission of the author.

There are 7 lexical chains in this example as follows:

- {first, initial, final}
- {night, dusk, darkness}
- {environment, setting, surrounding}
- {suburbs, driving, Volkswagen, car's, lights, windows, ceiling, commuters, traffic, Volks, apartment, city, suburbs, traffic, suburban, suburbs, residentialness, neighbourhood, community, suburban, drive, suburban, city, suburbs, apartments, Bloor St., Bathurst St., Finch St., driving, route, driven}
- {afflicted, darkness, panicky, mournful, exciting, deadly, hating, aversion, cruel, relentless, weird, eerie, cold, barren, sterile, hated, cruel, perversely}
- {married, wife, wife}

Chapter 3 explains how these chains are formed. Chapter 4 is an analysis of the correspondence of lexical chains to the structure of the text. The full example is given there, with details such as sentence numbers in which the words occur, and how the words are related.

1.3 Halliday and Hasan

The work of Halliday and Hasan [5] on cohesion has provided the foundation on which this thesis has developed. They provided a precise definition and a complete categorization of the five types of cohesion and the forms that each type can take. More important, their work has underscored the importance of cohesion as a good indicator of text unity. They introduced the idea of marking the sentences of a text with their cohesive properties, and using this information to help explain the text.

Halliday and Hasan's method of analyzing the cohesion in a text is different from that used in this thesis. They have categorized lexical cohesion as follows:

- same item {*bird, bird, bird*}
- synonym or near synonym, including hyponym {*feathered friend, bird*}
- superordinate {*element, stove*}
- general item {*that man*}
- collocation {*course, exam*}

For each of these types of lexical cohesion, the reference type may be one of the following:

- identical
- inclusive

- o exclusive
- o unrelated

This is illustrated in this example from Halliday and Hasan [5]:

(1-8) 1. There's a *boy* climbing that tree.
2. The *boy's* going to fall if he doesn't take care.
3. Those *boys* are always getting into mischief.
4. And there's another *boy* standing underneath.
5. Most *boys* love climbing trees.

As related to *boy* in sentence 1, sentence 2 has a reference type of ***identity***, sentence 3 is ***inclusive***, sentence 4 is ***exclusive***, and sentence 5 is ***unrelated***. This classification scheme is used in this thesis for the definition and recognition of lexical cohesion, but lexically cohesive words are not explicitly marked this way. The reason for this is that, in this thesis, lexical cohesion is used to indicate topical or coherent units of text, meaning that chains of lexical cohesion are significant regardless of the classification. It is expected that further research will include such a classification, as the type of cohesion used is related to the strength of the tie. An obvious example of this is the fact that the repetition of a word forms a very strong lexically cohesive chain.

Halliday and Hasan have employed a further classification of cohesion to codify the textual distance over which it is used. To them, distance is equal to the total number of sentences that occur between the text item that is presupposed in a cohesive tie, and the text item that presupposes it. A cohesive tie is considered as containing a presupposed text item and a text item that does the presupposing. (Note that the term *presupposing* does not mean presupposition in the normally used linguistic sense.) This is a precise definition for reference cohesion (and others), since in the typical case of a pronoun, the pronoun presupposes its referent. However, for lexical cohesion, the concept of a presupposed word is less precise. It may be a matter of interpretation as to whether two words are lexically tied. There does not have to be a cohesive tie as in the case of a pronoun for which there must be a referent.

According to their analysis, a cohesive tie can be immediate, mediated, or remote (also mediated *and* remote) depending on the distance relations between the presupposed item and the presupposing item. An immediate tie occurs when the two words are in consecutive sentences, or the same sentence. The distance is considered to be zero. A mediated tie occurs when a text item that presupposes another ties with a text item in the preceding sentence that depends on a previous text item for its ultimate resolution. An example of a mediated tie is:

(1-9) 1. John is going home.
2. He is finished work.
3. He worked twelve hours today.

The *He* in sentence 3 ties with the *He* in sentence 2 which depends on the *John* in sentence 1 for its ultimate resolution. The distance is 1, which is equal to the number of intermediate sentences in the chain of cohesion that contain an item that is both presupposed and presupposing. A remote cohesive tie occurs when there are sentences between the two cohesive items.

For the purpose of using lexically cohesive chains as indicators of coherent or topical chunks of text, one requires explicit lexical chains with a different distance measure, and an analysis of transitivity in the word relations. This will be discussed in chapters 3 and 4. Consider this example from *Alice in Wonderland* (offered by Halliday and Hasan):

(1-10)
1. The last word ended in a long bleat, so like a sheep that Alice quite started.
2. She looked at the Queen, who seemed to have suddenly wrapped herself up in wool.
3. Alice rubbed her eyes, and looked again.
4. She couldn't make out what had happened at all.
5. Was she in a shop?
6. And was that really—was it really a *sheep* that was sitting on the other side of the counter?
7. Rub as she would, she could make nothing more of it ....

They have marked the *wool* in sentence 2 to be an immediate lexically cohesive tie to the *sheep* in sentence 1. The *sheep* in sentence 6 is marked as a remote tie of distance 4 to the *sheep* in sentence 1.

1.4 Why Lexical Cohesion is Important

This section explains why lexical cohesion is important for computational text understanding systems. There are two major reasons:

1. Words must be interpreted in the context of related words. This aids in the resolution of ambiguity, and in the narrowing to a specific meaning of a word.

2. Lexical chains provide a clue for the determination of coherence and discourse structure, and hence the larger meaning of the text.

Furthermore, for the examples analysed in this thesis (4-1 to 4-5), the determination of lexical chains is a computationally feasible task.

1.4.1 Word Interpretation in Context

Meaning does not exist in isolation. Things and events are relative to other facts and experiences. Stated another way, texts are interpreted with respect to other related texts and situations that have occurred in the experience of the reader (Ventola [22], Halliday and Hasan [5], Phillips [15], Lancashire [12], Ide [11]). This context includes relevant or related portions of the current text.

Word meanings have this same property. They do not exist in isolation. They are interpreted with respect to other related words in a text. A system to understand text must therefore keep track of lexical chains for the following two related reasons:

1. By providing context, lexical chains allow for the narrowing to a specific word meaning.

2. Lexical chains provide a contextual clue to word sense disambiguation.

As an example of the first point, in the lexical chain {*gin, alcohol, sober, drinks*}, the meaning of *drinks* (*drinks* as a noun) is narrowed down to alcoholic *drinks*. To illustrate the second point, in the lexical chain {*hair, curl, comb, wave*} [5] *wave* means a hair wave, not a water wave, a physics wave, or a friendly hand wave. These facts about the meanings can be deduced with the aid of the lexical chain, and the chains, as the examples in this thesis indicate, are computationally feasible. It is important to note that the lexical chains do not stop at sentence boundaries. They can exist between a pair of adjacent words or range over an entire text.

Hirst [9] used a system called "Polaroid Words" to provide for intra-sentential lexical disambiguation. Polaroid Words are fake semantic objects that are given to the semantic interpreter called Absity in place of words. As information about the lexical sense of the word becomes available (unless there is no ambiguity) from communication with information about other words, the proper word sense information can be filled in. Hence the Polaroid Word will become a fully developed word. There is a Polaroid Word type for each syntactic category. It contains lexical knowledge about the word. For example, a Polaroid Word for a noun contains a list of the possible semantic objects that the noun could represent. This is an example from Hirst ([9], p. 103):

- slug (noun gastropod-without-shell bullet metal-stamping shot-of-liquor)

Polaroid Words communicate with each other by announcing a final choice or by sharing with "friends" the remaining available choices. Friends are considered to be the words that could possibly affect the lexical meaning of a word. As an example, friends of verbs are the prepositions and nouns that they dominate. So far, the theory of Polaroid Words has not been integrated with global context or discourse pragmatics. Hence it does not address the issue of word sense disambiguation across sentences.

1.4.2 Cohesion and Discourse Structure

The second major importance of lexical chains is that they provide a clue for the determination of coherence and discourse structure.

When a unit of text is about the same thing, there is a tendency for related words to be used. It follows that if lexical chains can be determined, they will tend to indicate the structure of the text. Furthermore, the examples analysed in this thesis indicate that the identification of lexical chains is a computationally feasible task, making it an important tool for text analysis in computational linguistics.

This section will describe the application of lexical cohesion to the determination of the discourse structure that was proposed by Grosz and Sidner [4]. First the theory will be briefly described, and then the relevance of lexical cohesion to it will be discussed.

In 1986, Grosz and Sidner developed a theory proposing a structure common to all discourse, which could be used along with a structurally dependent focus of attention to delineate and constrain referring expressions. In this theory there are three interacting components: *linguistic structure, intentional structure*, and *attentional state.*

Linguistic structure is the segmentation of discourse into groups of sentences, each fulfilling a distinct role in the discourse. Boundaries of segments are admittedly fuzzy, but some factors aiding in their determination are *clue words*, changes in intonation (not helpful in written text), and changes in aspect and tense. When found, these segments indicate changes in the topics or ideas being discussed, and hence will have an effect on what is referred to.

The second major component of the theory is the intentional structure. It is based on the idea that people have definite purposes for engaging in discourse. There is an overall discourse purpose, and also a discourse segment purpose for each of the segments in the linguistic structure described above. Each segment purpose specifies how the segment contributes to the overall discourse purpose.

There are two structural relationships between these segments. The first is called a *dominance* relation which occurs when the satisfaction (i.e., successful completion) of one segment's intention contributes to the satisfaction of another segment's intention. The second relation is called *satisfaction precedence*, which occurs when the satisfaction of one discourse segment purpose must occur before the satisfaction of another discourse segment purpose can occur.

The third component of this theory is the attentional state. This is a stack-based model of the set of things that attention is focused on at any given point in the discourse. It is "parasitic" on the intentional and linguistic structures, since for each discourse segment there exists a separate focus space. The dominance relations and satisfaction precedence relations determine the pushes and pops of this stack space. When a discourse segment purpose contributes to a discourse segment purpose of the immediately preceding discourse segment, the new focus space is pushed onto the stack. If the new discourse segment purpose contributes to a discourse segment purpose earlier in the discourse, focus spaces are popped off the stack until the discourse segment that the new one contributes to is on the top of the stack.

It is crucial to this theory that the linguistic segments be identified, and as stated by the authors, this is a problem area. This thesis shows that the lexical chains determined by the algorithm given in chapter 4 are a good indication of the linguistic segmentation. If a lexical chain ends, there is a tendency for a linguistic segment to end, since the lexical chains tend to indicate the topicality of segments. If a new lexical chain begins, this is an indication or clue that a new segment has begun. If an old chain is referred to again (this phenomenon is called *chain returns*), it is a strong indication that a previous segment is being returned to.

1.5 Cohesion and Coherence

The theory of *coherence relations* (Hobbs [10], Hirst [8], McKeown [14]) will now be considered in relation to cohesion. There has been some confusion as to the differences between the phenomena of *cohesion* and *coherence*. There is a danger of lumping cohesion and coherence together and losing the distinct contributions of each to the understanding of the unity of text.

Ultimately the difference between cohesion and coherence is this: cohesion is a term for sticking together; it means that the text all hangs together. Coherence is a term for making sense; it means that there is sense in the text. Hence *coherence relations* refers to the relations between sentences that allow them to make sense.

Cohesion and coherence relations may be distinguished in the following way. A coherence relation is a semantic dependency relation among clauses or sentences, such as *elaboration, support, cause,* or *exemplification.* There have been various attempts ([10], [14]) to classify all possible coherence relations, but there is as yet no widespread agreement. There does not exist a general computationally feasible mechanism for identifying these coherence relations.

In contrast, cohesion is defined as being one of the semantic dependency relations among elements in a text that were mentioned earlier: *reference, ellipsis, substitution, conjunction,* and *lexical cohesion.* There is a computationally feasible method for determining **lexical** cohesion, and this will be discussed at length in chapter 3 of this thesis.

Since cohesion is well-defined, one might expect that it would be computationally easier to identify, because the identity of the relation as one of *ellipsis, reference, substitution, conjunction,* and *lexical cohesion* is a straightforward task for people. This thesis will show that **lexical** cohesion is computationally feasible to identify. In contrast, the identification of a specific coherence relation from a given set is not a straightforward task for people such that the answer can be agreed upon. Consider this example from Hobbs [10] :

(1-11) 1. John can open Bill's safe.

2. He knows the combination.

Hobbs identifies the coherence relation as elaboration. I would call it explanation. This distinction depends on context, knowledge, and beliefs. For example, if you questioned *John's* ability to open *Bill's safe*, you would probably identify the relation as explanation. Otherwise you could identify the relation as elaboration. Here is another example:

(1-12) 1. John bought a raincoat.

2. He went shopping yesterday on Queen Street and it rained.

The coherence relation here could be elaboration (on the buying), or explanation (of when, and/or how, and/or why) or cause (he bought the raincoat because it was raining out).

The point is that the identity of coherence relations is "interpretative", whereas the identity of cohesion relations is not. At a general level, even if the precise coherence relation is not known, there exists the relation "is about the same thing" if coherence exists. Lexical cohesion is a strong contributor to this relation. In the example from Hobbs above, *safe* and *combination* are lexically related, which in a general sense means they "are about the same thing in some way". In example (1-12), *bought* and *shopping* are lexically related, as are *raincoat* and *rained.* This shows how cohesion can be useful in identifying sentences that are coherently related.

Cohesion and coherence are independent. Cohesion can exist in sentences that are not related coherently:

(1-13) 1. Wash and core six apples.

2. Use them to cut out the material for your new suit.

3. They tend to add a lot to the colour and texture of clothing.

4. Actually, maybe you should use five of them instead of six, since they are quite large.

Similarly, coherence can exist without cohesion:[2]

(1-14) 1. I came home from work at 6:00 p.m.

2. Dinner consisted of two chicken breasts and a bowl of rice.

It rarely happens however, that cohesion and coherence exclude each other. Most sentences that relate coherently exhibit cohesion as well. Example (1-14) is a list of sequential events, called a "time step" relation by Hirst [9]. This is one case where cohesion is not as prevalent, but where there is strong coherence.

There is an interesting analogy between cohesion and syntax, and coherence and semantics. *Jabberwocky* [2] is an example of syntax sticking text together without semantics. Example (1-13) illustrates cohesion sticking text together without semantics.

1.6 The Importance of Both Cohesion and Coherence

Halliday and Hasan give two examples of lexical cohesion ([5], p. 2,3) involving identity of reference:

(1-15) 1. Wash and core six cooking *apples*.

2. Put *them* into a fireproof dish.

(1-16) 1. Wash and core six cooking *apples*.

2. Put the *apples* into a fireproof dish.

Reichman ([17], p. 180) writes "It is not the use of a pronoun that gives **cohesion** to the wash-and-core-apples text. These utterances form a **coherent** piece of text not because the pronoun *them* is used but because they jointly describe a set of cooking instructions" (emphasis added). But this is wrong. Pronominal reference is defined as a type of **cohesion** (Halliday and Hasan [5]). Therefore the *them* in example one is certainly an instance of it (contradicting Reichman's first statement). The important point is that **both** cohesion and coherence are distinct phenomena creating unity in text.

Reichman also writes ([17], pp. 179) "that similar words (*apples, them, apples*) appear in a given stretch of discourse is an artifact of the content of discussion". It follows that if content is related in a stretch of discourse, there will be coherence. Lexical cohesion is a computationally feasible clue to identifying a coherent stretch of text. In example 1-16, it is computationally trivial to get the word relationship between *apples* and *apples*, and this relation fits the definition of lexical cohesion. Surely this simple indicator of coherence is useful, since as stated above, there does not exist a computationally feasible method of identifying coherence in non-domain-specific text.

Hobbs [10] sees the resolution of coreference (which is a form of cohesion) as being subsumed by the identification of coherence. He uses a formal definition of coherence

[2] Unless in the situation of knowing what normally happens after work, *dinner* is not considered to be cohesively related to *work*.

relations, an extensive knowledge base full of assertions and properties of objects and actions, and a mechanism that searches this knowledge source and makes simple inferences. Also, certain elements must be assumed to be coreferential.

He uses this example (the same as example 1-11):

(1-17) 1. John can open Bill's *safe*.

2. He knows the *combination*.

and shows how the *combination* gets identified as the combination of *Bill's safe* (he also shows how *John* and *He* are coreferential).

Lexical cohesion would be useful here to indicate that *safe* and *combination* can be assumed to be coreferential. But more importantly, one should not be misled by "what comes first, the chicken or the egg?" when dealing with cohesion and coherence. Rather, one should integrate the theories, using each phenomenon where applicable. Since the lexical cohesion between *combination* and *safe* is easy to compute, this thesis argues that it makes sense to use this information as an indicator of coherence.

1.7 Other Work on Lexical Cohesion in Text

1.7.1 Eija Ventola

Ventola [22] has analysed lexical cohesion and text structure within the framework of systemic linguistics and the domain of service encounters. An example of a service encounter is the exchange of words that takes place between a client at a post office and a postal worker. Within this framework, language is analysed in terms of registers and genres.

A *genre* is a type of text with common structural elements that reflect a similar global purpose and functionality. An example of a genre is the domain of Ventola's research, namely service encounters. Examples of structural elements common to this genre are greeting, service, pay, and closing.

A *register* consisting of three variables—field, mode, and tenor—captures the contextual or situational aspects and differences of a genre. *Field* contains the subject matter, purpose, and intent of the text. *Mode* refers to the medium of communication used, such as spoken or written, and to rhetorical mode such as narrative or didactic. *Tenor* refers to the role relations among the participants.

Ventola analysed service encounter text with respect to lexical cohesion to determine whether lexical cohesion reflected the registeral and generic structure outlined above. It was found that lexical cohesion provided an indication of certain generic elements of service encounters. More obviously, lexical cohesion reflected choices for field within the genre of service enounters such as postal matters or purchasing a travel ticket.

Ventola noted that one would expect a less marked effect of lexical cohesion in oral text, such as service encounters, since in oral text intonation can be used, and other semiotic codes are employed. Also, certain aspects of service encounters are so standardized that they need not be made explicit and language plays an "ancillary role".

An example (given by Ventola) of lexical strings reflecting generic structure is that in all service encounter texts a "rates string" was found. A "rates string" is a lexical string (called lexical chains in this thesis) containing words related to rates for service. The highest density of words in the "rates string" coincides with the structural element "pay". The structural element "pay" is common to all service encounter texts. Therefore, the lexical string is indicative of the structure of the text. When a new structural element starts, the lexical string that coincides with the old structural element stops.

An example of lexical cohesion reflecting register is that the string of words *padded postal bag, parcel,* and *tape* reflects the field of postal matters.

Although the domain and structural framework differ greatly from those used in this thesis, there is an important similarity: that of using lexical cohesion structures to explain and reflect global text structure. The genre of service encounters has a structure analogous to the structure of tying up a boat to a jetty [22]. There is a predictable sequence of functional events. This is in sharp contrast to the general-interest-short-article domain used in this thesis, where there is no analagous functional structure.

In Ventola's analysis of lexical cohesion, structure in the form of lexical strings is built. The string-building rule is that each lexical item is "taken back once to the nearest preceding lexically cohesive item regardless of distance"([22], p. 131), forming strings or chains of words that relate lexically. This methodology differs significantly from that used in this thesis, which is discussed in chapter 3.

In Ventola's work, transitivity of lexical relations is allowed to any level. A lexical string could be formed in the following way:

- word *a* is related to word *b*
- word *b* is related to word *c*
- word *c* is related to word *d*

and hence word *a* is related to word *d*. I believe that this will not produce strongly related lexical chains. The approach in this thesis is discussed in section 3.2.2.

1.7.2 Udo Hahn

Udo Hahn [6] has developed a text parsing system that considers both cohesion and coherence. The text (so far only nouns in the text) is mapped directly to the underlying knowledge model of the domain which is currently implemented as a frame-structured knowledge base.

Lexical cohesion is considered by Hahn to be a micro-structure phenomenon where local semantic relations occur between words in the text. Coherence is viewed as a global or macro structure of the text. One type of coherence structure used is patterns of regular thematic progression (Danes [3]), such as constant theme or linear thematization of rhemes. The *theme* is the topic or "given" part of the sentence, and the *rheme* is the comment or "new" part of the sentence. Consider the following sentence:

(1-18) 1. The boy is bad.

The theme is *The boy* and the rheme is *is bad.* Linear thematization of rhemes means making the rheme the theme in the next sentence. The other type of coherence structure used is special linguistically-marked relations such as comparison or contrast (not discussed further in Hahn's paper).

A major point of this work is that a correct analysis of textuality must include both cohesion and coherence analysis or the result will be understructured or misrepresented text.

As noted above, Hahn views lexical cohesion as a local phenomenon. Lexical cohesion is recognized between words in a sentence and the preceding sentence. There is an extended recognizer that works for cohesion contained within paragraph boundaries or on focus information. Focus information consists of the dominant concepts of each paragraph as determined through text condensation procedures (Hahn and Reimer [7]).

Recognizing lexical cohesion is a matter of searching for ways of relating frames and slots in the data base that are activated by words in the text. Nouns in the text activate a frame or slot when a match is found between them. Note that frames have the same name as certain pre-selected nouns, and these pre-selected nouns in the text correspond directly to frames in the data base. When a noun is encountered that corresponds to a data base frame, that frame is activated. If a noun corresponds to a slot or a slot value of an active frame, corresponding activation weights get increased. Hence, activation weights get increased each time a frame is used or a relation (twigged by nouns) is found between frames and slots. In Hahn's work, these activation weights are indicative of lexical cohesion. To illustrate this, consider the example from Hahn [6]:

(1-19)
1. The PC-1985 is equipped with a keyboard, a display, and a matrix printer of outstanding quality.
2. Moreover, that computer has a slightly less comfortable operating system, the common BASIC, and a functionally poor editor.

PC-1985 is a noun that correponds to a frame in the data base. Hence, the frame (with the name *PC-1985*) is activated. The word *keyboard* also corresponds to a frame in the data base. This frame is found to be a slot value for the peripheral devices slot of the frame *PC-1985*, and hence the frame named *keyboard* gets assigned as the slot value of the frame *PC-1985*, and the activation weight of the *keyboard* frame gets incremented.

Hahn considers two types of coherence relations: regular patterns of thematic progression, and other "special" linguistically marked relations like contrast or comparison. Lexical cohesion is used to reflect the coherence provided by thematic progression relations. An example of this is: if lexical cohesion relations activate a frame, and then a slot value of that frame gets activated in the next sentence, the result is a constant theme coherence relation. The "given" part of the sentence remains the same, and more is said about it in the next sentence.

In Hahn's work, heavy reliance is put on the "formally clear cut model of the underlying domain"([6], p. 3,4), which is made possible because of "far reaching constraints on the text propositions that are inherent in the semantic structure of the domain". However, general interest articles reflecting their author's opinions do not have domains that can be *a priori* formally represented as frames with slot values such that lexical cohesion and coherence will correspond directly to them. This becomes especially evident when considering the non-systematic semantic lexical relations.

1.7.3 Martin Phillips

Martin Phillips has developed a computer-aided, knowledge-free system to analyse lexical structures in text. The analysis in his work is based on a technique called distributed statistical analysis [15]. The results of the analysis are used to infer the "aboutness" or semantic meaning of texts. It is not a natural language understanding system in the normal sense (no world knowledge, parsing, syntax, or semantics), but rather a statistical technique for the analysis of text structure that is based on "collocation" patterns of non-function words in the text [13]. In Phillips's work, collocation refers to physical co-occurrence of words in a text within a set span. The span he used is four non-function words. Consider this example:

(1-20) 1. The parrot flew out the window.

In this example, *parrot* and *window* collocate since they are separated by less than four non-function words. Note that in this thesis, collocation is considered to be a semantic relation between words. Phillips uses distributed statistical analysis and cluster analysis [15] techniques to produce graphs or networks (of words and their collocational frequency patterns) that are intended to capture semantic content.

As stated above, the results of the statistical analysis are graphs. According to Phillips, analysis of these graphs reveals the presence of so called "central" words. Central words are those that have the most connections or links in the graph that represents the collocation patterns. He claims that the graph surrounding a central word narrows its meaning and allows new meanings to develop. Therefore, one of the major philosophies of his work is that a word's meaning is strongly related to the words that it co-occurs with in a text.

Phillips goes on to use these graphs to infer text macro-structure. He claims that the following is evidence for text macro-structure: three lexical graphs from two different chapters that can be superimposed with at least one common central word per superimposition. Furthermore, a text structure theory was developed based on how graphs representing chapters can link together. He used his analysis on five science text books and verified his structural results with the five authors of the books.

Phillips claims that the technique does not work well for literary text, since literary text (in his opinion) does not contain significant lexical structure. However, his lexical structures depend only on the frequency of physical co-occurrence of words, not on their semantic relationships. Since his work is not dependent on world knowledge such as lexicons, thesauri, or other data bases, it will not suffer the limitations imposed by creating and maintaining complete and up-to-date sources of knowledge.

Chapter 2

The Thesaurus and Lexical Cohesion

The thesaurus was first conceived by Peter Mark Roget in 1806. He envisioned a book where words would be classified according to the ideas they express. He finished it in 1852. In his introduction he described his thesaurus as being the "converse" of a dictionary. A dictionary explains the meaning of words, whereas a thesaurus, given an idea or meaning, aids in finding the words that best express the idea. For the creation of lexical chains, which are simply chains of words related by a common idea or meaning, thesaural knowledge will be useful.

2.1 The Structure of the Thesaurus

Roget's International Thesaurus (4th Edition) [19] is composed of 1042 basic sequentially numbered categories. There is a hierarchical structure both above and below this category level. Figure 2.1 is an example of this structure for category 407, which is labelled life.

There are three structure levels above the category level. The top-most level consists of eight major *classes* developed by Roget in 1852. These eight classes are: abstract relations, space, physics, matter, sensation, intellect, volition, and affections.

Each class is divided into (roman-numbered) *subclasses.* For example, under Class 4, matter, there are three subclasses as illustrated below:

Class 4: Matter
  I. Matter in General
  II. Inorganic Matter
  III. Organic Matter

Under each subclass there is a (capital-letter-sequenced) *sub-subclass.* For example in Class 4, matter, subclass III, organic matter, there are six sub-subclasses as follows:

III. Organic Matter
  A. Animal and Vegetable Kingdom
  B. Vitality
  C. Vegetable Life

Figure 2.1: The structure of Roget's Thesaurus

Class 1 ···

⋮

Class 4: Matter

  I ···

  ⋮

  III Organic Matter

    A ···

    ⋮

    B Vitality

      ⋮

      407 Life

        1. NOUNS life, living, vitality, being alive, having life, animation, animate existence; liveliness, animal spirits, vivacity, spriteliness; long life, longevity; viability; lifetime 110.5; immortality 112.3; birth 167; existence 1; bio -, organ -; -biosis.

          ⋮

        2. ···

        ⋮

      408 Death ···

      ⋮

    ⋮

  ⋮

⋮

D. Animal Life
E. Mankind
F. Male and Female

Below the semantic category level, there are *syntactic categories* consisting of: noun, verb, adverb, preposition, conjunction, interjection, and "phrases". Phrases is a catch-all for related expressions. There are also sequentially numbered *paragraphs* below the category level for closely related words within a category. Within a paragraph there are *semi-colon groups* of more closely related words (the ";" is used as the group marker). There are also *cross-references* or pointers to other categories in the thesaurus that are related to the current semi-colon group. These pointers can be either a category number such as 407, or a category number and paragraph number such as 407.1.

Where applicable, categories are organized into *antonym pairs*. As an example, category 407 is *Life*, and category 408 is *Death*.

The thesaurus contains an index, which allows for a quick lookup of words that are related to a given word. For each word, the index contains a list of word-labelled category numbers. The category numbers can be either a category, or a category and paragraph. These categories in the list contain words that are related to the index word. As an example, consider the index entry for the word *lid*:

Lid

clothing	231.35
cover	228.5
eyelid	439.9
stopper	266.4

The index is important for this thesis, since it is used by the thesaurus lookup methods that determine lexical relations. This process is discussed in depth in section 3.3.

2.2 Differences from Traditional Knowledge Bases

In traditional artificial intelligence knowledge bases such as frames or semantic networks, words or ideas that are related are actually physically close in the representation. In a thesaurus this need not be true. Physical closeness has some importance, as can be seen clearly from the hierarchy described above (in section 2.1), but words in the index of the thesaurus often have widely scattered categories, and each category often points to a widely scattered selection of categories. As an example, consider the index entry for the word *lid* given in section 2.1. Often the index entries of the thesaurus have categories that range over most of the thesaurus.

The thesaurus simply groups words by idea. It does not have to name or classify the idea or relationship. In traditional computer databases, the relationship must be named. For example in a semantic net, a relationship might be **isa** or **colour-of**, and in a frame database, there might be a slot for **colour** or **location**.

In chapter 1, different types of word relationships were discussed: systematic semantic, non-systematic semantic, word association, and words related by a common situation. A common factor to all but situational relationships is that there is a strong tendency for the word relationships to be captured in the thesaurus. This holds even for the non-systematic semantic relations, which are the most problematic by definition. A thesaurus simply groups related words without attempting to explicitly name each relationship. In a traditional computer database, a systematic semantic relationship can be represented by a slot value for a frame, or by a named link in a semantic network. If it is hard to classify a relationship in a systematic semantic way, it will be hard to represent the relationship in a traditional frame or semantic network formalism. Of the 16 non-systematic semantic lexical chains given as examples in Halliday and Hasan [5], 14 were found in *Roget's Thesaurus* [18]. This represents an 87% hit rate (but not a big sample space). Word associations show a strong tendency to be findable in a thesaurus. Of the 16 word association pairs given in [9], 14 were found in *Roget's Thesaurus* [18]. Since two of the word senses were not contained in the thesaurus, this represents a 100% hit rate among those that were. Situational word relationships are not as likely to be found in a general thesaurus.

2.3 Other Work on Thesauri in Text Understanding

Sedelow and Sedelow [21] have done a significant amount of research on the thesaurus (in particular Roget's) as a valuable representation of knowledge for use in a natural language understanding system.

Their work has concentrated on the lower-level structures of the thesaurus (like semicolon groups) to "obviate some of the difficulties which may be latent in the Aristotelian and Enlightenment sort of scheme for structuring knowledge which Roget used". They point out that problems with the upper-level structure have misled many researchers into concluding that the thesaurus is not a valid knowledge source for text understanding. They also point out the importance of general non-domain-specific approaches to text understanding, and the fact that the point of a thesaurus is to contain general non-domain-specific semantic relations between words.

In most work on measures of close semantic distance or relatedness, it is physical closeness of a knowledge structure that is used to determine the strength of a relationship. The system of Polaroid Words (see section 1.4.1) is a good example of this. Sedelow and Sedelow emphasize that although that type of closeness does indeed imply a close semantic relation, this should not imply that physically far-removed entries cannot be closely related semantically. In fact both physically close and far-removed entries are closely related in a thesaurus, which helps to explain its large benefits in dealing with the non-systematic semantic relations that are normally problematic.

Sedelow and Sedelow have been interested in the application of clustering patterns in the thesaurus to natural language understanding. One application used the idea that if two words sharing the same stem (where at least one of the words has a prefix) are found in the same or nearby sections of the thesaurus, this should be a good clue to identifying the function of the prefixes. For example, *prevent* was analysed as non-prefixed since the word *prevent* doesn't occur in categories with the stem *vent*.

Robert Bryan [1] has proposed a graph-theoretic model of the thesaurus. A boolean matrix is created with words on one axis and categories on the other. A cell is marked as true if a word associated with a cell intersects with the category associated with a cell. Paths or chains in this model are formed by travelling along rows or columns to other true cells. Semantic "neighbourhoods" are grown, consisting of the set of chains emanating from an entry. It was found that without some concept of chain strength, the semantic relatedness of these neighbourhoods decays, partially due to homographs. Strong links are defined in terms of the degree of overlap between categories and words. A strong link exists where at least two categories contain more than one word in common, or at least two words contain more than one category in common. The use of strong links was found to enable the growth of strong semantic chains with homograph disambiguation. Consider this example matrix, where the columns are categories, and the rows are words:

	c1	c2	c3	c4
w1	T	F	T	F
w2	F	F	F	F
w3	T	F	T	T

In this example, there is a strong link between categories 1 and 3 since they contain two words in common. There is also a strong link between words 1 and 3, since they contain two categories in common. There is not however, a strong link between categories 3 and 4 since they contain only only one word in common.

This concept is different from that used in this thesis. Here, by virtue of words co-occurring in a text and then also containing at least one category in common or being in the same category, they are considered lexically related and no further strength is needed. I use the thesaurus as a validator of lexical relations that are possible due to the semantic relations among words in a text.

Chapter 3

Finding Lexical Chains

3.1 General Methodology

This thesis describes a text-understanding tool that builds lexical chains, and uses them as an aid in determining the structure of the text. This chapter details how these lexical chains are formed, using a thesaurus as the main knowledge base. The tool is intended to be useful for text that is not domain-specific. This has not been the emphasis of computational linguistics in the past because of the unmanageable computational complexity involved. Here, lexical cohesion is viewed as a general phenomenon that is computable. There are five major examples presented in full detail in chapter 4. These examples consist of (sometimes parts of) general-interest articles from five magazines: *New Yorker*, *Reader's Digest*, *Equinox*, *Toronto*, and the *Life* section of the *Toronto Star*.

There are lexical chains existing in these examples that a person can find using common sense and intuition. But the aim of this chapter is to describe a method of automating the computation of them. The process used to accomplish this for the five examples is as follow:

1. Identify intuitive chains using common sense and a knowledge of English.
2. Find values for the following parameters necessary to compute the intuitive chains:
   - thesaural relations
   - transitivity of word relations
   - distance (in sentences) between words in a chain
3. Formalize the results of step 2, for all five examples, into a general algorithm.

It must be kept in mind that this is intended to be a computationally feasible system. The aim was to find efficient, intuitively plausible methods that will cover enough cases to ensure the production of meaningful results. Note that in this thesis, "intuitive" means the result of using both common sense and a knowledge of English. Thesaural lookup methods were sought and obtained that ensure that there is a legitimate lexical relation. These thesaural relations must not relate all words meaninglessly. Transitivity for the lexical chain relations was found that allowed the intuitive chains to be computable, but that does not result in interference from words that should not be considered members of a chain.

The process described above was done by hand. Automation was not possible because of a lack of an online thesaurus with the lookup methods required. It would be easy to design such a computer system, since it would require only traditional data base search and lookup techniques that have been in existence for years. It is expected that further research involving an automated system run on a large example space would give valuable information on the fine-tuning of the parameter settings used in the general algorithm.

3.2 Forming Lexical Chains

3.2.1 Candidate Words

The first decision in lexical chain formation is which words in the text are chain candidates. As pointed out by Halliday and Hasan [5], repetitive occurrences of closed-system words such as pronouns, prepositions, and verbal auxiliaries are obviously not considered. Also, high frequency words like *good, do*, and *taking* do not normally enter into lexical chains with some exceptions such as *takings* used in the sense of earnings. As an example, consider the first two sentences of example (4-2):

(3-1) 1. My *maternal grandfather lived* to be *111.*

2. *Zayde* was *lucid* to the end, but a few years before he *died* the *family assigned* me the *task* of talking to him about his *problem* with *alcohol.*

Only the italicised words were considered as lexical chain candidates.

3.2.2 Building Chains

3.2.2.1 Thesaural Relations

Once the candidate words are chosen, the lexical chains can be formed. The major knowledge base used for chain computation was the thesaurus. In this work an abridged version of *Roget's Thesaurus* [18] was used. Five types of thesaural relations were found to be necessary, but the first two types were by far the most prevalent, validating over 90% of the lexical relationships. Section 2.1 gives an explanation of the structure of the thesaurus, including the index, categories, pointers, and labels. The following are the thesaural relationships used to form the lexical chains:

1. Two words have a category in their index entries in common. As an example (from example (4-1) chain 1), *residentialness* and *apartment* both have category 189 in their index entries. A pictorial representation of this relation is given in figure 3.1 (a).

2. One word has a category in its index entry that contains a pointer to a category of the other word. As an example (from example (4-1) chain 1) *car* has category 273 in its index entry that contains a pointer to category 276, which is a category of the word *driving.* A pictorial representation of this relation is given in figure 3.1 (b).

3. A word is either a label in the other word's index entry, or is in a category of the other word. Note that an index entry contains labelled category numbers (see section 2.1). As an example (from example (4-2) chain 11), *blind* has category 442 in its index entry, which contains the word *see*. A pictorial representation of this relation is given in figure 3.1 (c).

4. There is a structural relation where words are in category pairs meaning that they are antonyms, or they are in the same group, and hence semantically related. As an example (from example (4-2) chain 11) , *blind* has category 442, blindness, in its index entry and *see* has category 441, vision, in its index entry. A pictorial representation of this relation is given in figure 3.1 (d).

5. The two words have categories in their index entries that both point to a common category. For example (from example (4-5) chain 1), *brutal* has category 851 that has a pointer to category 830. *Terrified* has category 860 that has a pointer to category 830. A pictorial representation of this relation is given in figure 3.1 (e).

Figure 3.1: Thesaural Relations

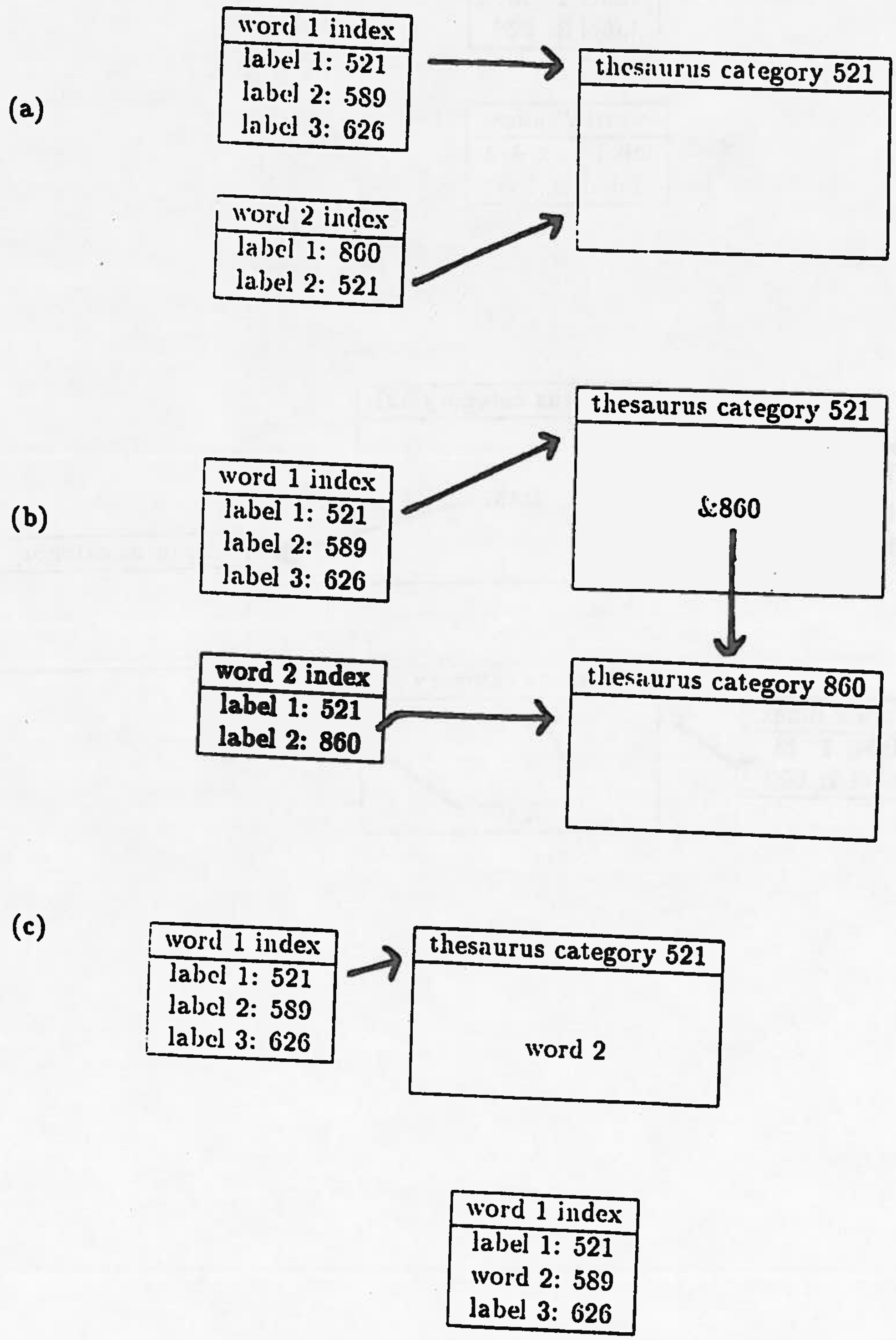
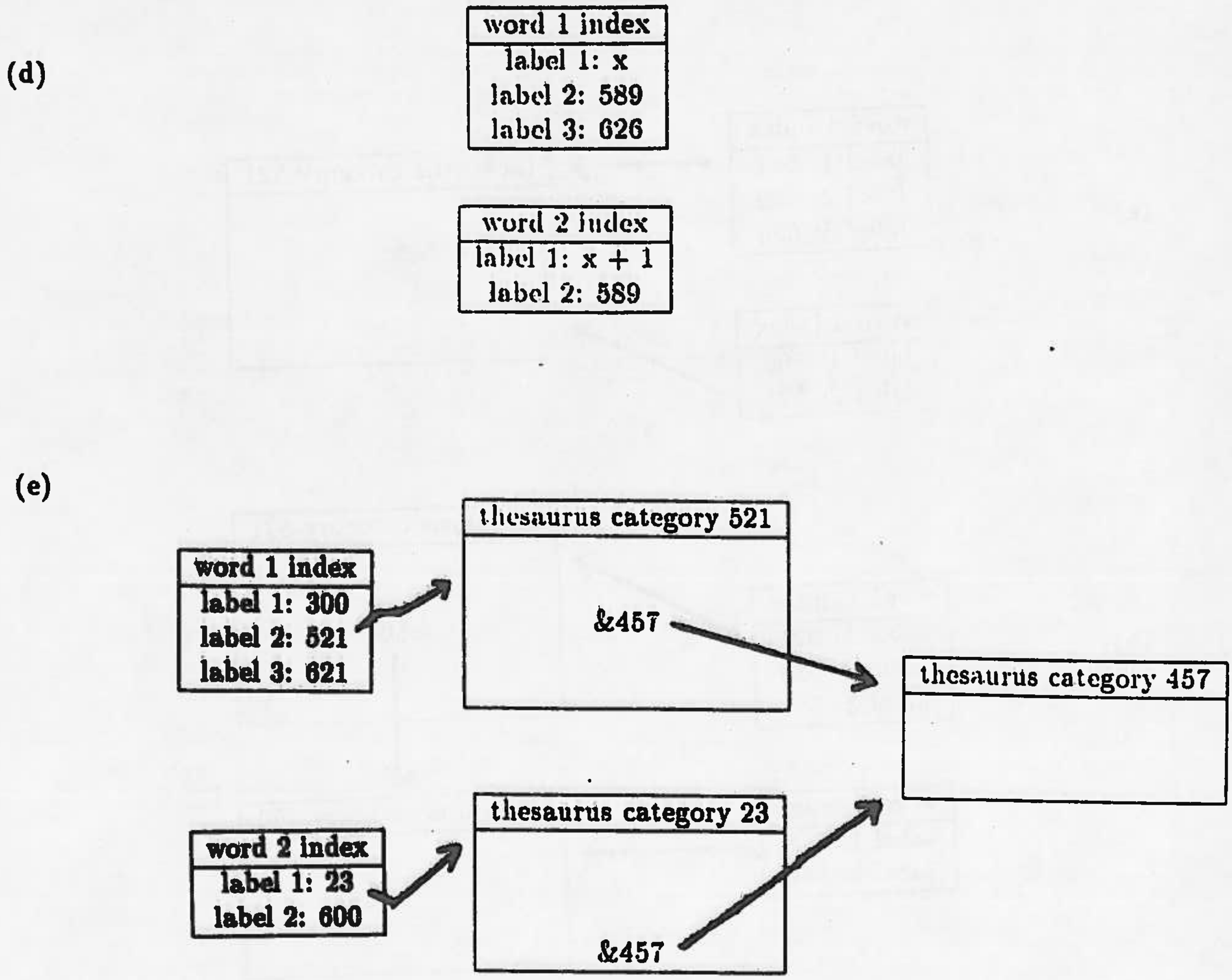


Figure 3.1 Continued



All of the five examples given in this thesis (in chapter 4) used mostly thesaural relation types 1 and 2. Examples (4-2) to (4-5) also used type 5 rarely (less than 5% of the time), example (4-2) also used type 3 rarely, and example (4-4) also used type 4 rarely.

3.2.2.2 Transitivity Relations Used

When computing lexical chains, the question of how much transitivity to allow arises. Specifically, if:

- word *a* is related to word *b*
- word *b* is related to word *c*
- word *c* is related to word *d*

then is word *a* related to words *c* and *d*?

My intuition was to allow one transitive link. In the above example this means that word *a* is related to word *c* but not to word *d*. It seemed that two or more transitive links would so severely weaken the word relationship as to cause it to be non-intuitive. Consider this chain: {*cow, sheep, wool, scarf, boots, hat, snow*}. If unlimited transitivity were allowed, then *cow* and *snow* would be considered related which is definitely counter-intuitive.

There are two ways in which a transitive relation involving one link can cause two words to be related. They are shown in figure 3.2. In type one, if *word1* is related to *word2*, and *word2* is related to *word3*, then this implies that *word1* is related to *word3*. In type two, if *word1* is related to *word2*, and *word1* is related to *word3*, then this implies that *word2* is related to *word3*. Lexical chains are calculated only with respect to the text read so far. For example, if *word5* is related to *word3* and *word5* is related to *word4*, then *word3* and *word4* are not related, since at the time of processing, words 3 and 4 were not relatable.

Since one transitive link was viewed as the intuitively correct approach, the examples had to be analysed in order to answer the question: does one transitive link allow for all of the intuitive chains to be computed, especially as an aid if there is no direct thesaural link between the words?

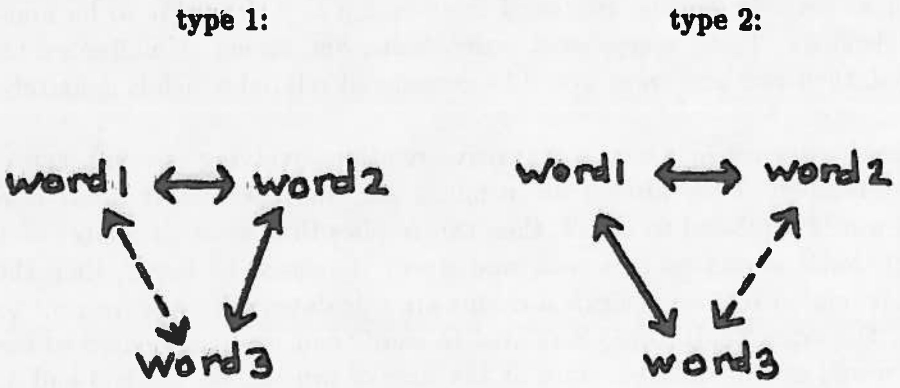
In examples (4-1), (4-2), (4-4), and (4-5), the answer to this question is yes. Even unlimited transitivity would not have improved lexical chain computation. In example (4-3), chain 2, a transitivity of two would have allowed all words except *lack* and *strange* to enter into one chain. In fact, I had separated the chain into two intuitive chains with {*rudely, strange, failing, lack, afflicted, bad*} forming a separate chain, but the thesaurus caused the unfortunate links.

To summarize, a transitivity of one link is sufficient to successfully compute the intuitive chains. An automated system could be used to test this out extensively, varying the number of transitive links and calculating the consequences. It is likely that it varies slightly with repect to style, author or type of text.

3.2.2.3 Distance Between Words in a Chain and Chain Returns

We now consider how many sentences can separate two words in a lexical chain before the words are considered to be unrelated. Related to this question is the phenomenon of

Figure 3.2: Transitive Relations



type 1:

type 2:

chain returns. Sometimes, several sentences after a chain has clearly stopped, the chain will be returned to. It would have been easier to simply let a new chain start, and not try to relate new chains back to existing ones to form returns. However, returns are used to link together larger expanses of text than are contained in single chains or chain segments. Returns to existing chains often correspond to intentional boundaries since they occur after digressions or sub-intentions (see section 1.4.2 for an explanation of intentional structure), signalling a resumption of some structural text entity. Each part of a chain containing returns is called a *chain segment.*

It seems intuitive that the distance between words in a chain is a factor in chain formation. It also seems obvious that the distance will not be "large", because words in a text co-relate due to recognizable relations, and large distances would interfere with the recognition of relations. Note that the sentence was chosen here as the unit of distance, but that other units such as the word or clause could be used.

The five examples were analysed with respect to distance between words. The analysis showed that there can be up to two or three intermediary sentences between a word and a chain with which it can be linked. For distances of four or more intermediary sentences, the word is only able to signal a return to an existing chain. It was found that more than two or three intermediary sentences can exist between a chain and a return to it. In the five examples used in this thesis, returns happened after between four and 19 intermediary sentences. One significant fact emerged from this analysis: returns consisting of one word only were always made with a repetition of one of the words in the returned-to chain. Returns consisting of more than one word did not necessarily use repetition, in fact in most cases, the first word in the return was not a repetition.

The question of chain returns and when they can occur requires further research. When distances between relatable words are not tightly bounded (as in the case of returns) the chances of unfortunate unintuitive chain linkages increases. It is anticipated that chain return analysis would become integrated with other text processing tools in order to prevent this. Also, I believe that chain **strength** analysis will be required for this purpose. It is possible that only strong chains can be returned to. Chain strength and factors affecting it are discussed in the next section.

3.2.3 Chain Strength

It seems intuitive that some lexical chains are "stronger" than others. There are three factors contributing to chain strength:

1. Reiteration—the more, the stronger.
2. Density—the denser, the stronger.
3. Closeness of words in the chain—the closer, the stronger.

Ideally, some combination of values reflecting these three factors should result in a chain strength value that can be useful in determining if a chain is strong enough to be returned to. Also, a strong chain should be more likely to have a structural correspondence than a weak one. It seems likely that chains could contain particularly strong portions with special implications for structure. These issues will not be addressed here.

3.2.4 Notation and Data Structures

The information stored for each word in each chain includes the following:

- A word number, which is a sequential, chain-based number for each word so that it can be uniquely identified.
- The sentence number in which the word occurs.
- The chain created so far.

Each lexical relationship in a chain is represented as $(u, v)_x^y$ where:

- u is the current word number
- v is the word number of the related word
- x is the transitive distance:
  - 0 means no transitive link was used to form the word relationship
  - 1 means one transitive link was used to form the word relationship
- y is either
  - the thesaural relationship number given in section 3.2.2.1
  - Tq where
    * T stands for transitively related
    * q is the word number through which the transitive relation is formed.

A full example of this notation is taken from example (4-2) chain 9:

Chain 1		
Word	Sentence	Lexical Chain
1. evade	15	
2. feigning	15	$(2, 1)_0^2$
3. escaped	16	$(3, 1)_1^1$ $(3, 2)_1^{T2}$

Word 2, *feigning*, is lexically related to word 1, *evade*, through thesaural relation type number 2, and there is no transitive link used to form the relationship. *Evade* has category 477 in its index entry which contains a pointer to category 544. *Feign* has category 544 in its index entry. Word 3, *escaped*, is related to word 1, *evade*, through thesaural relation type 1, and there is no transitive link used. *Evade* has category 927 in its index entry , and *escape* also has category 927 in its index entry. Also at word 3, calculation of the chain so far using a transitivity of one link means that word 3 is related to word 2 transitively through word 1.

Figure 3.3: Algorithm for Finding Lexical Chains

```
REPEAT
        READ next word
        IF word is suitable for lexical analysis (see section 3.2.1) THEN
                CHECK for chains within a suitable span
                (up to 3 intermediary sentences, and no limitation on
                returns):
                        CHECK thesaurus in a suitable manner (section 3.2.2.1).
                        CHECK other knowledge sources
                        (situational, general words, proper names).
                IF chain relationship is found THEN
                        INCLUDE word in chain.
                        CALCULATE chain so far
                        (allow one transitive link).
                END IF
                IF there are words that have not formed a chain for a suitable
                number of sentences (up to 3) THEN
                        ELIMINATE words from the span.
                END IF
                CHECK new word for relevance to existing chains that
                are suitable for checking.
                ELIMINATE chains that are not suitable for checking.
        END IF
END REPEAT
```

3.2.5 General Algorithm

Figure 3.3 shows the generalized algorithm for computing lexical chains. Parameter values are given in brackets for the following:

- candidate words
- thesaural relations
- transitivity of word relations
- distance between words in a chain

The parameter values were determined from an analysis of the five examples given in chapter 4, and are explained in sections 3.3.1 and 3.3.2. The only parameter not addressed in this thesis is which (if any) chains should be eliminated from the chain-finding process.

3.3 Problems and Concerns

This section is a discussion of problems encountered during the computation of the lexical chains contained in the five examples given and analysed in chapter 4.

One problem encountered was that occasionally the algorithm would cause two chains to merge together, whereas intuition would lead one to keep them separate. In example (4-1), Chain 1, I had found the following separate chain beginning in sentence 38: {*people, Metropolitan Toronto, people, urban, population, people, population, population, people*}. However, the algorithm linked this chain in with Chain 1 (that runs through the entire example) consisting of these words and others: {*city, suburbs, traffic, community*}. Fortunately, this was a rare occurrence. Note that there will be cases where the lexical chains should be merged. This will happen as a result of the intentional merging of ideas or concepts in the text.

Another issue that came up was whether or not to use lexical *rendering*. This is a term used by Ventola [22] that means resolving reference cohesion, and including this information in the lexical chains. For example, pronouns could then enter into lexical chains. This has its own set of problems however. Consider example (4-2), Chain 2, which is a chain containing family member words such as {*family, aunt*}. Words 12 and 13 are both *grandfather*, which I had intuitively chosen as a separate chain. The words *Zayde* and especially *he* are used repetitively throughout the entire article, and if the pronouns were used in the chain it would run throughout the entire example and the separate chain of words 12 and 13 would be lost.

Another problem occurs when words form relations according to the algorithm but no intuitive relation exists. In example (4-4), chain 8, word 5 is *courses* related thesaurally to the word *political* from chain 6. This happened because homograph disambiguation is not handled, and the other non-academic sense of the word *course* came into play. If the chains were used to disambiguate homographs this would not occur. Fortunately, this problem is rare (in the examples analysed in chapter 4).

There is one general problem that should be mentioned. The algorithm described in this chapter is not automated, and was run on five examples of real-life text to get a feel for values for various parameters (transitivity, thesaural relations, distance). Research with an automated system would allow large amounts of text to be analysed, giving more reliable parameter settings. Also, an automated system could be used to track variations of parameters over style, author, and type of text [1].

3.3.1 Where the Thesaurus Failed as a Validator of Lexical Relations

The thesaurus validated well over 90% of the lexical relations from the five examples in this thesis. The following is an example-by-example analysis of when the thesaurus failed to validate a relationship and why.

In example (4-1), chain 6, the intuitive chain {*hand-in-hand*, matching, whispering, *laughing, warm*} was not entirely computable. Only the italicised words were relatable. The words in chain 6 are cohesive by virtue of being general, but strong, "good" words related by their goodness, rather than by their specific meanings. Chain 10, {*environment,*

[1] This idea was suggested by Eduard Hovy during a discussion of this work.

setting, surrounding}, was not thesaurally relatable. *Setting* was not in the thesaurus, and it seems as though *environment* and *surrounding* should be thesaurally connected, but they were not.

In example (4-2), chain 12, word 4, *blurted*, does not relate thesaurally to the rest of the chain which contains {*uneasily, trouble, hurt, afraid, excited*}. The missing knowledge (the knowledge not contained in the thesaurus) is **situational** knowledge. One *blurts* rather than *says* something under the conditions alluded to in chain 12.

In example (4-3), chain 2 there were a couple of validation problems. The chain {*searched, surveyed, bother, tear gas*} did not relate thesaurally to the rest of chain segments 2.1 and 2.2. Chain 2.3 is not considered to be a return to chain 2. For an explanation of chain returns and segments see section 3.2.2.3. The knowledge that is missing here is situational, the situation being that of security checks. *Searching, surveying, tear gas,* and being *bothered* are a part of security checks. In chain segment 2.3 (considered as separate from the rest of chain 2), the words {*bad, afflicted, rudely*} did not relate thesaurally to the words {*disruptive, strange, failing, insecurity, lack*}. The intuitive chain 2.3 is made up of words connoting generally "bad" things which explains why the thesaurus failed to relate them. The thesaurus groups words by specific meaning relations, not by general "goodness" or "badness" qualities.

In example (4-4), chain 6, word 2, *law* did not relate to the chain that contained {*rights, power, official, policy*}. This is puzzling since this is the kind of relationship expected to be contained in a thesaurus. The thesaurus did not relate all of chain 7. Instead these three chains were created: {*full-time, full-time, full-time*}, {*jobs, work, work, job, working, work, support, job, work, worker, job*}, and {*salary, money, salaries, wage*}. Obviously these three chains are intuitively related. The missing knowledge is situational. We know that *full-time* is a particular *working* situation and that *work* is normally carried out for the purpose of making *money.* In chain 8, the words *dissertation* and *papers* did not relate thesaurally to the chain containing {*educated, academic, graduate, study, courses, exams, college*}. Again, the missing knowledge is knowledge of a particular situation, in this case the "academic" situation.

In example (4-5), chain 3, which formed one intuitive chain, was split by the thesaurus into three parts: the words *innocent* and *executioner* that did not relate to anything, {*sentenced, confessed, implicate, accused, arrests*}, and {*court, interrogation, search, denied*}. Once again, it is a lack of situational knowledge in the thesaurus that is causing the problem. The situation here is that of the court process. Chain 5 is another example of the lack of situational knowledge in the thesaurus that causes it to fail to relate words. Three chains were created using the thesaurus whereas there is only one intuitive chain. The chains are {*rats, rats, rats*}, {*corruption, filth, squalid, fleas*}, and {*poor, poverty, pest*} whose words all form one intuitive chain by virtue of being semantically related or associated with the situation of *poverty* and *squalor.* Chain 7 is yet another example of the lack of situational knowledge in the thesaurus. The words {*homosexual, prostitute, intravenous, drugs*} are related by being common to the situation of having *AIDS*, but they are not relatable using the thesaurus. Chain 8 exemplifies another type of knowledge that the thesaurus does not provide. The chain is {*moral, divine, evangelists, Christian*} and the word that does not thesaurally relate to this chain is *St. Paul's.*

Place names, street names, and people's names are generally not to be found in *Roget's*

Thesaurus [18]. However, they are certainly contained in one's "mental thesaurus". Example (4-1), chain 1, which contains several major Toronto street names, is another good example of this. These names are certainly related to the rest of chain 1 in my mental thesaurus since I am currently a resident of Toronto.

To summarize, there were few cases where the thesaurus failed to validate an intuitive lexical chain. For those cases where the thesaurus used did fail, there are three missing knowledge sources that became apparent:

1. General semantic relations between words that are strongly "good" or "bad" (or "something").
2. Situational knowledge.
3. Specific proper names like city or street names.

All of this knowledge is embodied in the "mental" thesaurus that enables one to form intuitive lexical chains from a text.

3.3.2 Problems with Distances and Chain Returns

There were few exceptions to the rule allowing a maximum of three intermediary sentences between words forming a chain. In example (4-2), chain 2, word 5, *family*, relates by distance rules to word 6 which is also *family*, but this is counter-intuitive. In Chain 12, word 2 in sentence 25 relates intuitively to word 3 in sentence 30, but the distance rule disallows this. This brings up the point that sentence *length* must also be considered. In this case, the four intermediary sentences are very short, thus enabling the intuitive relationship. This happens again in chain 16, where word 1 in sentence 28 relates intuitively to word 2 in sentence 33, and the four sentences in between are very short. Also, in this case, the *strike one* sets up a structural expectation for *strike two*.

In example (4-5), chain 1, which runs steadily throughout the entire example, there was one case where there were four intermediary sentences between words (words 8 and 9). It does not make sense in this case to have two separate chains simply because in one instance there were four intermediary sentences.

There were a few cases of unfortunate chain returns occurring where they were definitely counter-intuitive. In example (4-1), chain 3, word 4, *wife* is not considered as part of the rest of the chain {*married, wife, wife*}. It would be a one-word return to chain 3; however there is no intuitive reason to link them. This is simply a case of an unfortunate chain return.

In example (4-2), chain 11, I did not initially believe that segment 11.2 was an intuitive return to chain 11.1. After analysing the example a great deal, I could see that maybe it is an intuitive return. It is the only case in the five examples from chapter 4 that is ambiguous. I have decided to stick with my first reaction and consider segment 11.2 as an unfortunate and counter-intuitive chain return. Text understanding is an interpretive process at some level, and so an automated text understanding system should not be expected to come up with the right interpretation in all cases, since what is right is debatable. Also, lexical structure analysis is used in this thesis as a tool for providing clues to a text structure analyser. Therefore it is used more as a first cut at structure determination, and not as a producer of deep and complete structural analysis.

In example (4-3), chain 2.3 related to words in chain segments 2.1, and 2.2, and hence would have formed a chain return, not a new chain. There are differences between the chains that common sense can distinguish, but a thesaurus cannot. When forming an intuitive lexical chain, the chain meaning is considered. Chain meaning is the collective meaning resulting from all of the individual word meanings. Chain segments 2.1, and 2.2 are about the negative aspects of security, whereas chain segment 2.3 is about other bad things, specifically the inability of the Bolshoi Ballet to generate rapport with its audience.

In example (4-4), chain 8, the word *academic* was repeated twice in sentence 13, forming a chain. Then, from sentences 27 to 31, there is a chain of eight words relating to graduate study. The initial part of chain 8 should not be related to the second part. Perhaps chain strength analysis could help here, since segment 8.1 is not a strong chain, and should not be returned to.

Chapter 4

How to Use the Lexical Chains

This chapter describes how the lexical chains (formed by using the algorithm given in chapter 3) can be used as a tool in an automated text understanding system. As outlined in section 1.3, there are two areas of text understanding where lexical chains are useful:

- determining text structure
- word interpretation in context

This chapter considers only the application of lexical chains to text structure analysis.

4.1 Lexical Chains and Text Structure

Any structural theory of text must be concerned with identifying units of text that are about the same thing. When a unit of text is about the same thing there is a strong tendency for semantically related words to be used within that unit. By definition, lexical chains are chains of semantically related words. Therefore it makes sense to find them and use them as an aid to determining the structure of the text. This is particularly true if it is feasible to compute the chains. For the examples analysed in this chapter, the lexical chains are computable using the algorithm in section 3.2.5.

This section will concentrate on analysing correspondences between lexical chains and structural units of text including:

- the correspondence of chain boundaries to structural unit boundaries
- returns to existing chains and what this indicates about structural units
- lexical chain strength and reliability of predicting correspondences between chains and structural units
- an analysis of problems encountered and when extra textual information is required to validate the correspondences between lexical chains and structural components

The text structure theory chosen for this analysis is the intentional theory proposed by Grosz and Sidner [4]. This theory is described in section 1.4.2. It was chosen for several reasons. Firstly, it is an attempt at a general non-domain-dependent theory of text structure, and that is the domain of this work as indicated by the five general-interest examples chosen for analysis. Secondly, it has gained a significant acceptance in the field as a good standard approach.[1] Thirdly, it is relatively easy to understand how the theory works and hence to apply it to new text examples.

The methodology used in the following five analyses is as follows:

1. Determine the lexical chain structure of the text using the algorithm given in section 3.2.5. In certain rare cases where the algorithm does not form intuitive lexical chains properly, it is noted, both in section 3.3 and in the analysis in this chapter. The intuitive chain is used for the analysis, however the lexical chain data given in this chapter will show the rare mismatches between intuition and the algorithm.

2. Determine the *intentional structure* of the text using the theory outlined by Grosz and Sidner [4]. The structure produced from the application of their theory is called the intentional structure and the structural components are called *intentions.*

3. Compare the lexical structure formed in step 1 with the intentional structure formed in step 2, and analyse for correspondences between them.

Once each example has been analysed in this way, overall conclusions can be reached, and problems identified.

4.2 Example (4-1)

4.2.1 The Text

Here is the text of example (4-1), the first section of an article in *Toronto* magazine, December 1987, by Jay Teitel, entitled "Outland":[2]

(4-1)

1. ¶I spent the first 19 years of my life in the suburbs, the initial 14 or so relatively contented, the last four or five wanting mainly to be elsewhere.
2. The final two I remember vividly: I passed them driving to and from the University of Toronto in a red 1962 Volkswagen 1500 afflicted with night blindness.
3. The car's lights never worked—every dusk turned into a kind of medieval race against darkness, a panicky, mournful rush north, away from everything I knew was exciting, toward everything I knew was deadly.
4. I remember looking through the windows at the commuters mired in traffic beside me and actively hating them for their passivity.
5. I actually punched holes in the white vinyl ceiling of the Volks and then, by way of penance, wrote beside them the names and phone numbers of the girls I would call when I had my own apartment in the city.
6. One thing I swore to myself: I would never live in the suburbs again.
7. ¶My aversion was as much a matter of environment as it was traffic—one particular piece of the suburban setting: the "cruel sun."

[1] Robin Cohen, personal communication.

[2] ©1987 Jay Teitel. Reprinted by kind permission of the author.

8. Growing up in the suburbs you can get used to a surprising number of things—the relentless "residentialness" of your surroundings, the weird certainty you have that everything will stay vaguely new-looking and immune to historic soul no matter how many years pass.

9. You don't notice the eerie silence that descends each weekday when every sound is drained out of your neighbourhood along with all the people who've gone to work.

10. I got used to pizza, and cars, and the fact that the cultural hub of my community was the collective TV set.

11. But once a week I would step outside as dusk was about to fall and be absolutely bowled over by the setting sun, slanting huge and cold across the untreed front lawns, reminding me not just how barren and sterile, but how undefended life could be.

12. As much as I hated the suburban drive to school, I wanted to get away from the cruel suburban sun.

13. ¶When I was married a few years later, my attitude hadn't changed.

14. My wife was a city girl herself, and although her reaction to the suburbs was less intense than mine, we lived in a series of apartments safely straddling Bloor Street.

15. But four years ago, we had a second child, and simultaneously the school my wife taught at moved to Bathurst Street north of Finch Avenue.

16. She was now driving 45 minutes north to work every morning, along a route that was perversely identical to the one I'd driven in college.

17. ¶We started looking for a house.

18. Our first limit was St. Clair—we would go no farther north.

19. When we took a closer look at the price tags in the area though, we conceded that maybe we'd have to go to Eglinton—but that was definitely it.

20. But the streets whose names had once been magical barriers, latitudes of tolerance, quickly changed to something else as the Sundays passed.

21. Eglinton became Lawrence, which became Wilson, which became Sheppard.

22. One wind-swept day in May I found myself sitting in a town-house development north of Steeles Avenue called Shakespeare Estates.

23. It wasn't until we stepped outside, and the sun, blazing unopposed over a country club, smacked me in the eyes, that I came to.

24. It was the cruel sun.

25. We got into the car and drove back to the Danforth and porches as fast as we could, grateful to have been reprieved.

26. ¶And then one Sunday in June I drove north alone.

27. This time I drove up Bathurst past my wife's new school, hit Steeles, and kept going, beyond Centre Street and past Highway 7 as well.

28. I passed farms, a man selling lobsters out of his trunk on the shoulder of the road, a chronic care hospital, a country club and what looked like a mosque.

29. I reached a light and turned right.

30. I saw a sign that said Houses and turned right again.

31. ¶In front of me lay a virgin crescent cut out of pine bush.

32. A dozen houses were going up, in various stages of construction, surrounded by hummocks of dry earth and stands of precariously tall trees nude halfway up their trunks.

33. They were the kind of trees you might see in the mountains.

34. A couple was walking hand-in-hand up the dusty dirt roadway, wearing matching blue track suits.

35. On a "front lawn" beyond them, several little girls with hair exactly the same colour of blond as my daughter's were whispering and laughing together.

36. The air smelled of sawdust and sun.

37. ¶It was a suburb, but somehow different from any suburb I knew.

38. It felt warm.

39. ¶It was Casa Drive.

40. ¶In 1976 there were 2,124,291 people in Metropolitan Toronto, an area bordered by Steeles Avenue to the north, Etobicoke Creek on the west, and the Rouge River to the east.

41. In 1986, the same area contained 2,192,721 people, an increase of 3 percent, all but negligible on an urban scale.

42. In the same span of time the three outlying regions stretching across the top of Metro—Peel, Durham, and York —increased in population by 55 percent, from 814,000 to some 1,262,000.

43. Half a million people had poured into the cresent north of Toronto in the space of a decade, during which time the population of the City of Toronto actually declined as did the populations of the "old" suburbs with the exception of Etobicoke and Scarborough.

44. If the sprawling agglomeration of people known as Toronto has boomed in the past 10 years it has boomed outside the traditional city confines in a totally new city, a new suburbia containing one and a quarter million people.

4.2.2 The Lexical Structure

The following tables show the lexical chains found in example (4-1):

Chain 1		
Word	Sentence	Lexical Chain
1. suburbs	1	
2. driving	2	
3. Volkswagen	2	
4. car's	3	$(4, 2)_0^2$
5. lights	3	
6. commuters	4	
7. traffic	4	$(7, 2)_0^2$ $(7, 4)_0^1$
8. Volks	5	
9. apartment	5	$(9, 1)_0^1$
10. city	5	$(10, 1)_0^1$ $(10, 2)_0^1$ $(10, 4)_0^{T2}$ $(10, 7)_0^1$ $(10, 9)_0^1$
11. suburbs	6	$(11, 1)_0^0$ $(11, 9\text{-}10)_0^1$ $(11, 2\text{-}7)_1^{T10}$
12. traffic	7	$(12, 2)_0^2$ $(12, 4\text{-}10)_0^1$ $(12, 7)_0^0$ $(12, 11)_2^{T10}$
13. suburban	7	$(13, 1\text{-}11)_0^0$ $(13, 9\text{-}10)_0^1$ $(13, 2\text{-}12)_1^{T10}$
14. suburbs	8	$(14, 1\text{-}11\text{-}13)_0^0$ $(14, 9\text{-}10\text{-}13)_0^1$ $(14, 2\text{-}12)_1^{T10}$
15. residentialness	8	$(15, 1\text{-}9\text{-}10\text{-}13\text{-}14)_0^1$ $(15, 2\text{-}7\text{-}12)_1^{T10}$
16. neighbourhood	9	$(16, 1\text{-}11\text{-}13\text{-}14)_0^1$ $(16, 9\text{-}10\text{-}13)_1^{T14}$
17. community	10	
18. suburban	12	$(18, 1\text{-}11\text{-}13\text{-}14)_0^0$ $(18, 9\text{-}10\text{-}16)_0^1$ $(18, 2\text{-}12)_1^{T10}$
19. drive	12	$(19, 2)_0^0$ $(19, 7\text{-}10\text{-}12)_0^1$ $(19, 4)_0^2$ $(19, 1\text{-}9\text{-}11\text{-}13\text{-}14\text{-}15\text{-}16\text{-}18)_1^{T10}$ $(20, 9\text{-}10\text{-}16)_0^1$ $(20, 2\text{-}12\text{-}19)_2^{T10}$
20. suburban	12	$(20, 1\text{-}11\text{-}13\text{-}14\text{-}18)_0^0$ $(20, 9\text{-}10\text{-}16)_0^1$ $(20, 2\text{-}12\text{-}19)_1^{T1}$
21. city	14	$(21, 10)_0^0$ $(21, 1\text{-}2\text{-}7\text{-}9\text{-}13\text{-}14\text{-}15\text{-}16\text{-}19)_1^{T10}$ $(21, 4\text{-}12)_1^{T19}$
22. suburbs	14	$(22, 1\text{-}11\text{-}13\text{-}14\text{-}18\text{-}20)_0^0$ $(22, 9\text{-}10\text{-}16\text{-}21)_0^1$ $(22, 2\text{-}12\text{-}19)_1^{T10}$
23. apartments	14	$(23, 9)_0^0$ $(23, 1\text{-}10\text{-}11\text{-}13\text{-}14\text{-}15\text{-}16\text{-}18\text{-}20\text{-}21\text{-}22)_0^1$ $(23, 2\text{-}4\text{-}7\text{-}12\text{-}19)_1^{T21}$

Chain 1 (continued)		
Word	Sentence	Lexical Chain
24. Bloor St.	14	
25. Bathurst St.	15	
26. Finch St.	15	
27. driving	16	$(27, 2-19)^0_0$ $(27, 7-10-12-21)^1_0$ $(27, 4)^2_0$ $(27, 1-9-11-13-14-15-16-18-20-22-23)^{T10}_2$
28. route	16	$(28, 1-2-9-10-11-13-14-15-16-18-19-20-21-22-23-27)^2_0$ $(28, 4-7-12)^{T27}_1$
29. driven	16	$(29, 2-19-27-29)^0_0$ $(29, 7-10-12-21)^1_0$ $(29, 4-28)^2_0$ $(29, 1-9-11-13-14-15-16-18-20-22-23)^{T10}_1$
30. house	17	$(30, 1-9-10-11-13-14-15-16-18-20-21-22-23)^1_0$ $(30, 2-4-7-12-19-27-28-29)^{T10}_1$
31. St. Clair	18	
32. Eglinton	19	
33. streets	20	$(33, 1-10-13-14-15-16-18-20-21-22-23-30)^1_0$ $(33, 2-4-7-12-19-27-28-29)^{T10}_1$
34. Eglinton	21	
35. Lawrence	21	
36. Wilson	21	
37. Sheppard	21	
38. town-house	22	$(38, 30)^0_0$ $(38, 1-10-13-14-15-16-18-20-21-22-23)^1_0$ $(38, 2-4-7-12-19-27-28-29-33)^{T10}_1$
39. Steeles	22	
40. car	25	$(40, 2-19-27-29)^1_0$ $(40, 4-7-10-12-21-28)^{T29}_1$
41. drove	25	$(41, 2-19-27-29)^0_0$ $(41, 7-10-12-21)^1_0$ $(41, 4-28)^2_0$ $(41, 1-9-11-13-14-15-16-18-20-22-30-38)^{T10}_1$
42. Danforth	25	
43. porches	25	$(43, 33)^1_0$ $(43, 1-4-10-13-14-15-18-20-21-22-23-30-38-40)^2_0$ $(43, 16)^{T38}_1$ $(43, 2-19-23-29)^{T40}_1$
44. drove	26	$(44, 2-19-27-29-41)^0_0$ $(44, 7-10-12-21)^1_0$ $(44, 4-28)^2_0$ $(44, 1-9-11-13-14-15-16-18-20-22-23-30-38)^{T10}_1$
45. drove	27	$(45, 2-19-27-29-41-44)^0_0$ $(45, 7-10-12-21)^1_0$ $(45, 4-28)^2_0$ $(45, 1-9-11-13-14-15-16-18-20-22-23-30-38)^{T10}_1$
46. Bathurst	27	

Chain 1 (continued)		
Word	Sentence	Lexical Chain
47. Steeles	27	
48. Centre St.	27	
49. Highway 7	27	
50. trunk	28	
51. road	28	(51, 1-9-10-11-13-14-15-16-18-20-21-22-23-28-30-38)$^{1}_{0}$ (51, 43)$^{2}_{0}$ (51, 7)$^{T10}_{1}$ (51, 16)$^{T38}_{1}$
52. light	29	(52, 5)$^{0}_{0}$
53. turned	29	
54. houses	30	(54, 30-38)$^{0}_{0}$ (54, 1-9-10-11-13-14-15-18-20-21-22-23-33-43-52)$^{1}_{0}$ (54, 16-28)$^{2}_{0}$ (54, 2-7-12-19-29-41-44)$^{T10}_{1}$
55. turned	30	(55, 53)$^{0}_{0}$
56. houses	32	(56, 30-38-54)$^{0}_{0}$ (56, 1-9-10-11-13-14-15-18-20-21-22 -23-33-43-51)$^{1}_{0}$ (56, 16-28)$^{2}_{0}$ (56, 2-7-12-19-29-41-44)$^{T10}_{1}$
57. roadway	34	(57, 51)$^{0}_{0}$ (57, 1-9-10-11-13-14-15-16-18-20-21-22-23 -28-30-38)$^{1}_{0}$ (57, 43)$^{2}_{0}$ (57, 7)$^{T10}_{1}$ (57, 16)$^{T38}_{1}$
58. lawn	35	(58, 1-9-10-11-13-14-15-18-20-21-22-23-30-33-38-43-51-54-56-57)$^{1}_{0}$ (58, 28)$^{5}_{0}$ (58, 2-12-19-27-29-41-44)$^{T10}_{1}$ (58, 16)$^{T56}_{1}$
59. suburb	37	(59, 1-11-13-14-18-20-22)$^{0}_{0}$ (59, 30-38-56)$^{1}_{0}$ (59, 9-10-15-21-23-33-43-51)$^{1}_{0}$ (59, 16-28)$^{2}_{0}$ (59, 2-7-12-19 -29-41-44)$^{T10}_{1}$
60. suburb	37	(60, 1-11-13-14-18-20-22-59)$^{0}_{0}$ (60, 30-38-56)$^{1}_{0}$ (60, 9-10-15-21-23-33-43-51-54-56-57-59)$^{1}_{0}$ (60, 16-28)$^{2}_{0}$ (60, 2-7-12-19-29-41-4446-47)$^{T10}_{1}$
61. people	40	(61, 15)$^{1}_{0}$ (61, 1-9-10-11-13-14-18-20-21-22-23-30-33-38-51-54-56-57-59-60)$^{2}_{0}$ (61, 2-7-12-19-27-29-41-44)$^{T10}_{1}$ (61, 16-43-58)$^{T56}_{1}$
62. Metropolitan Toronto	40	(62, 1-9-10-11-13-14-15-18-20-21-22-23-30-33-38-51 -54-56-57-59-60)$^{1}_{0}$ (62, 2-7-12-19-27-29-41-44)$^{T10}_{1}$ (62, 16-43-58)$^{2}_{0}$
63. Steeles	40	
64. people	41	(64, 61)$^{0}_{0}$ (64, 15)$^{1}_{0}$ (64, 1-9-10-11-13-14-18-20-21-22-23-30-33-38-51-54-56-57-59-60-62)$^{2}_{0}$ (65, 2-7-12-19-27-29-41-44)$^{T10}_{1}$ (61, 16-43-58)$^{T56}_{1}$

Chain 1 (continued)		
Word	Sentence	Lexical Chain
65. urban	41	$(65, 1-9-10-11-13-14-15-18-20-21-22-23-30-33-38-51-54-56-57-59-60-62)_0^1$ $(65, 2-7-12-19-27-29-41-44)_1^{T10}$ $(65, 16-43-58)_0^2$
66. Metro	42	$(66, 62)_0^0$ $(66, 1-9-10-11-13-14-15-18-20-21-22-23-30-33-38-51-54-56-57-59-60)_0^1$ $(66, 2-7-12-19-27-29-41-44)_1^{T10}$ $(66, 16-43-58-64)_0^2$
67. Peel	42	
68. Durham	42	
69. York	42	
70. population	42	$(70, 30-38-54-56-61-64)_0^1$ $(70, 1-9-10-11-13-14-15-18-20-21-22-23-33-51-57-59-60-62-65-66)_0^2$ $(70, 43-58)_0^5$ $(70, 2-7-12-19-27-29-41-44)_1^{T10}$ $(70, 16)_1^{T64}$
71. people	43	$(71, 61-64)_0^0$ $(71, 15-70)_0^1$ $(71, 1-9-10-11-13-14-18-20-21-22-23-30-33-38-51-54-56-57-59-60-62-65-66)_0^2$ $(71, 2-7-12-19-27-29-41-44)_1^{T10}$ $(71, 16-43-58-64)_1^{T56}$
72. Toronto	43	
73. population	43	$(73, 70)_0^0$ $(73, 30-38-51-54-56-61-65-71)_0^1$ $(73, 1-9-10-11-13-14-15-18-20-21-22-23-33-51-57-59-60-62-65-66)_0^2$ $(73, 43-58)_0^5$ $(73, 2-7-12-19-27-29-41-44)_1^{T10}$ $(73, 16)_1^{T64}$
74. city	43	$(74, 10-21)_0^0$ $(74, 1-2-7-9-11-12-13-14-15-18-19-20-22-23-27-29-30-33-38-41-44-51-54-56-57-59-60-62-65)_0^1$ $(74, 16-28-43-58-65-70-71-73)_0^2$ $(74, 4-40)_1^{T47}$
75. Toronto	43	
76. population	43	$(76, 70-73)_0^0$ $(76, 30-38-54-56-61-64-71)_0^1$ $(76, 1-9-10-11-13-14-15-18-20-21-22-23-33-51-57-59-60-62-65-66-74)_0^2$ $(76, 43-58)_0^5$ $(76, 2-7-12-19-27-29-41-44)_1^{T10}$ $(76, 16)_1^{T64}$
77. suburbs	43	$(77, 1-11-13-14-18-20-22-59-60)_0^0$ $(77, 30-38-56-62-65-66-74)_0^1$ $(77, 9-10-15-21-23-33-43-51)_0^1$ $(77, 16-28-64-70-71-72-73-76)_0^2$ $(77, 2-7-12-19-29-41-44-)_1^{T10}$

Chain 1 (continued)		
Word	Sentence	Lexical Chain
78. Etobicoke	43	
79. Scarborough	43	
80. people	44	(80, 61-64-71)$^0_0$ (80, 15-70)$^1_0$ (80, 1-9-10-11-13-14-18-20-21-22-23-30-33-38-51-54-56-57-59-60-62-65-66-73-76-77)$^2_0$ (80, 2-7-12-19-27-29-41-44)$^{T10}_1$ (80, 16-43-58)$^{T56}_1$
81. Toronto	44	
82. city	44	(82, 10-21-74)$^0_0$ (82, 1-2-7-9-11-12-13-14-15-18-19-20-22-23-27-29-30-33-38-41-44-46-47-51-54-56-57-59-60-62-65-77)$^1_0$ (82, 16-28-43-58-64-70-71-73-76-80)$^2_0$ (82, 4-40)$^{T47}_1$
83. suburbia	44	(83, 1-11-13-14-18-20-22-59-60-77)$^0_0$ (83, 30-38-56-82)$^1_0$ (83, 9-10-15-21-23-33-43-51-82)$^1_0$ (83, 16-28-80)$^2_0$ (83, 2-7-12-19-29-41-44)$^{T10}_1$
84. people	44	(84, 61-64-71-80)$^0_0$ (84, 15-70-82)$^1_0$ (84, 1-9-10-11-13-14-18-20-21-22-23-30-33-38-51-54-56-57-59-60-62-65-66-73-76-77-82)$^2_0$ (84, 2-7-12-19-27-29-41-44)$^{T10}_1$ (84, 16-43-58)$^{T56}_1$

Chain 2, Segment 1		
Word	Sentence	Lexical Chain
1. afflicted	2	
2. darkness	3	$(2, 1)_0^2$
3. panicky	3	$(3, 1)_0^2$ $(3, 2)_0^5$
4. mournful	3	$(4, 1)_0^1$ $(4, 2)_0^1$ $(4, 3)_0^2$
5. exciting	3	$(5, 1\text{-}4)_0^2$ $(5, 2\text{-}3)_0^5$
6. deadly	3	$(6, 1\text{-}4)_0^2$ $(6, 2\text{-}3\text{-}5)_0^5$
7. hating	4	$(7, 1\text{-}4)_0^1$ $(7, 2\text{-}3\text{-}5\text{-}6)_0^2$
8. aversion	7	$(8, 7)_0^1$ $(8, 1\text{-}4)_0^2$ $(8, 2\text{-}3\text{-}5\text{-}6)_0^5$
9. cruel	7	$(9, 1\text{-}4\text{-}7)_0^1$ $(9, 2\text{-}3\text{-}5\text{-}6\text{-}8)_0^2$
10. relentless	8	$(10, 9)_0^1$ $(10, 1\text{-}4\text{-}7)_0^2$ $(10, 2\text{-}3\text{-}5\text{-}6\text{-}8)_0^5$
11. weird	8	$(11, 3)_0^1$ $(11, 1\text{-}4\text{-}7\text{-}10)_0^2$ $(11, 2\text{-}3\text{-}5\text{-}6\text{-}8)_0^5$
12. eerie	9	$(12, 3\text{-}11)_0^1$ $(12, 1\text{-}4\text{-}7\text{-}10)_0^2$ $(12, 2\text{-}3\text{-}5\text{-}6\text{-}8)_0^5$
13. cold	11	$(13, 3\text{-}6\text{-}7\text{-}8\text{-}11\text{-}12)_0^1$ $(13, 1\text{-}4\text{-}9)_0^2$ $(13, 2\text{-}3\text{-}5\text{-}6\text{-}10)_0^5$
14. barren	11	$(14, 6\text{-}7)_0^2$ $(14, 1\text{-}2\text{-}3\text{-}4\text{-}5\text{-}8\text{-}9\text{-}10\text{-}11\text{-}12\text{-}13)_1^{T7}$
15. sterile	11	$(15, 14)_0^1$ $(15, 6\text{-}7)_0^2$ $(15, 1\text{-}2\text{-}3\text{-}4\text{-}5\text{-}8\text{-}9\text{-}10\text{-}11\text{-}12\text{-}13)_1^{T7}$
16. hated	12	$(16, 7)_0^0$ $(16, 1\text{-}4\text{-}6\text{-}8\text{-}9\text{-}13)_0^1$ $(16, 14\text{-}15)_0^2$ $(16, 2\text{-}3\text{-}5\text{-}10\text{-}11\text{-}12)_0^5$
17. cruel	12	$(17, 9)_0^0$ $(17, 1\text{-}4\text{-}7\text{-}10)_0^1$ $(17, 2\text{-}3\text{-}5\text{-}6\text{-}8\text{-}11\text{-}12\text{-}13)_0^5$ $(17, 14\text{-}15)_1^{T7}$

Chain 2, Segment 2		
Word	Sentence	Lexical Chain
18. perversely	16	$(18, 10)_0^2$ $(18, 1\text{-}2\text{-}3\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}11\text{-}12\text{-}13\text{-}16\text{-}17)_1^{T10}$

Chain 2, Segment 3		
Word	Sentence	Lexical Chain
19. cruel	24	$(19, 9\text{-}17)_0^0$ $(19, 1\text{-}4\text{-}7\text{-}10)_0^1$ $(19, 2\text{-}3\text{-}5\text{-}6\text{-}8\text{-}11\text{-}12\text{-}13)_0^5$ $(19, 14\text{-}15)_1^{T7}$

Chain 3		
Word	Sentence	Lexical Chain
1. married	13	
2. wife	14	$(2, 1)^{1}_{0}$
3. wife	15	$(3, 1)^{1}_{0}$ $(3, 2)^{0}_{0}$
4. wife	27	$(4, 2\text{-}3)^{0}_{0}$ $(4, 1)^{1}_{0}$

Chain 4		
Word	Sentence	Lexical Chain
1. conceded	19	
2. tolerance	20	$(2, 1)^{1}_{0}$

Chain 5		
Word	Sentence	Lexical Chain
1. virgin	31	
2. pine	31	
3. bush	31	$(3, 1)^{1}_{0}$
4. trees	32	$(4, 1)^{1}_{0}$ $(4, 3)^{1}_{0}$
5. trunks	32	
6. trees	33	$(6, 4)^{0}_{0}$ $(6, 1\text{-}3)^{1}_{0}$

Chain 6		
Word	Sentence	Lexical Chain
1. hand-in-hand	34	
2. matching	34	
3. whispering	35	
4. laughing	35	
5. warm	38	$(5, 1)^{1}_{0}$ $(5, 4)^{5}_{0}$

Chain 7		
Word	Sentence	Lexical Chain
1. first	1	
2. initial	1	$(2, 1)^1_0$
3. final	2	$(3, 2\text{-}1)^3_0$

Chain 8		
Word	Sentence	Lexical Chain
1. night	2	
2. dusk	3	$(2, 1)^2_0$
3. darkness	3	$(3, 1\text{-}2)^1_0$

Chain 9		
Word	Sentence	Lexical Chain
1. environment	7	
2. setting	7	
3. surrounding	8	

4.2.3 The Intentional Structure

Figure 4.1 gives the intentional structure for this example.

Figure 4.1: The Intentional Structure of Example (4-1)

1 (1–44)
Describe changing attitudes to suburban life.

1.1 (1–25)
Describe earlier aversion to suburban life.

1.1.1 (1–7)
Describe hatred of commuting.

1.1.2 (8–12)
Describe the hated suburb environment.

1.1.3 (13–25)
Describe how this old aversion to suburbs held, when a recent attempt was made to buy a new house in the suburbs.

1.1.3.1 (13–16)
Describe how life changed, giving author reason to look for a new house.

1.1.3.2 (17–22)
Describe that houses are too expensive in Metro Toronto, hence one must look in the suburbs to buy a house.

1.1.3.3 (23–25)
Describe how the old familiar aversion to suburbs came back.

1.2 (26–39)
Describe a new suburb that seems livable in and nice.

1.2.1 (26–30)
Describe the drive to the new suburb.

1.2.2 (31–33)
Describe the bush-like area.

1.2.3 (34–39)
Describe the pleasant environment.

1.3 (40–44)
Describe why the new suburbs exist.

4.2.4 The Correspondences between Lexical and Intentional Structures

The following table gives the correspondences between the lexical chains and intentions of example (4-1):

Chain	Chain Range	Intention	Intention Range
1	1–44	1	1–44
2.1	2–12	1.1.1,1.1.2	1–12
2.2	16	end of 1.1.3.1	16
2.3	24	end of 1.1.3.3	25
3	13–15	1.1.3.1	13–16
4	19–20	1.1.3.2	17–22
5	31–33	1.2.2	31–33
6	34–38	1.2.3	34–39
7,8	1–3	1.1.1	1–7
9	7–8	1.1.2	8–12

There is a clear correspondence between chain 1 and intention 1. The continuity of the subject matter is reflected by the continuous lexical chain. From sentences 40 to 44, two words, *population* and *people* are used repetitively in the chain. *Population* is reiterated three times, and *people* is reiterated five times. If chain strength (indicated by the reiteration) were used to delineate "strong" portions of a chain, this strength information could be used to indicate structural attributes of the text. Specifically, sentences 40 to 44 form intention 1.3, and hence a strong portion of the chain would correspond exactly to a structural unit. *Drive* was repeated eight times between sentences 2 to 26, corresponding to intention 1.1. *Suburb* was repeated eleven times throughout the entire example indicating the continuity in structure between sentences 1 to 44.

Chain 2.1, from sentences 2 to 12, corresponds to intentions 1.1.1 and 1.1.2. More textual information is needed in order to separate intentions 1.1.1 and 1.1.2. There is a one word return to chain 2 at sentences 16 and 24, strongly indicating that chain 2 corresponds to intention 1.1, which runs from sentences 1 to 25. Also, segment 2.2 coincides with the end of intention 1.1.3.1, and segment 2.3 coincides with the end of intention 1.1.3.3. This situation illustrates why chain return analysis is necessary. Remember that after processing two to three sentences containing no words in the chain, the chain is considered to end. However, if a new chain is started, a check is made with existing chains to see if the new chain is a return to an existing chain. If chain returns were not considered, chain 2 would end at sentence 12, and the structural implications of the two single word returns would be lost. It is intuitive that the two words *perverse* and *cruel* indicate links back to the rest of intention 1.1. The link provided by the last return, *cruel*, is especially strong since it occurs after the diversion describing the attempt to find a nice house in the suburbs. *Cruel* is the third reiteration of the word in chain 2.

Chain 3 corresponds to intention 1.1.3.1. This is an example of an unfortunate chain return. Word 4, *wife*, would be considered a chain return by the algorithm, but is not an intuitive return, and so it is not considered to be a return here. If it were, the return would be coincident with the end of intention 1.1.3, but again, since this was not intuitive, it is not considered significant. Note that, as stated earlier, both common sense, and a knowledge of English combine to create intuition.

Chain 4 corresponds to intention 1.1.3.2, and the boundaries of the lexical chain are two sentences inside the boundaries of the intention. The existence of a lexical chain is a clue to the existence of a separate intention. Boundaries within one or two sentences of the intention boundaries are considered to be close matches.

Chain 5 corresponds closely to intention 1.2.2. Chain 6 corresponds closely to intention 1.2.3. Chains 7 and 8 are a couple of short chains (three words long) that overlap. They collectively correspond to intention 1.1.1. The fact that they are short and overlapping suggests that they could be taken together as a whole. This is a case where other information such as intentional knowledge, coherence relations, or semantics should be used validate the correspondence.

Chain 9 corresponds to intention 1.1.2. Even though the chain is a lot shorter in length than the intention, its presence is a clue to the existence of a separate intention in its textual vicinity. Since the lexical chain boundary is more than two sentences away from the intention boundary, other textual information would be required to validate the correspondence.

Overall, the lexical chains found in this example provide a good clue for the determination of the intentional structure. In some cases, the chains correspond exactly to an intention. It should also be stressed, however, that the lexical structures cannot be used on their own to predict an exact structural partitioning of the text. This of course was never expected. As a good example of the limitations of the tool, intention 1.2 starts in sentence 26, but there are no new lexical chains starting there. The only clue to the start of the new intention would be the ending of chain 2.

This example provides a good illustration (chain 2) of the importance of chain returns being used to indicate a high level intention spanning the length of the entire chain (including all segments). Also, the returns coincided with intentional boundaries.

4.3 Example (4-2)

4.3.1 The Text

Here is the text of example (4-2), an article in the Canadian edition of *Reader's Digest* magazine, December 1987, by A.M. Clarfield, entitled "A Grandson's Mission":[3]

(4-2) 1. ¶My maternal grandfather lived to be 111!

2. Zayde was lucid to the end, but a few years before he died, the family assigned me the task of talking to him about his "problem" with alcohol.

3. ¶My aunt, with whom he had lived for 20 years, was worried about my grandfather's desire to indulge, three to four times a day, in a drink of his favourite whiskey, fretting that he was about to become an alcoholic.

[3] Copyright © 1987 A.M. Clarfield. Reprinted by kind permission of Dr A.M. Clarfield and Reader's Digest Magazines Limited.

4. ¶He could not understand her fears and would slip himself a few drinks above and beyond the watered-down ration she would dole out each evening before supper.
5. I was a medical student at the time, and because I represented the closest thing to medical authority, I was delegated to “ speak to Zayde” about his tippling.
6. ¶Not exactly brimming with enthusiasm, I walked the few blocks to his house.
7. Climbing the 20 steep stairs to his room, I pondered how I should broach the delicate subject.
8. ¶After all, I had great respect for my Zayde.
9. He had come to Canada penniless, devoid of all but three English words (“I vant vork”) and had made it—bringing up a family and starting a hardware shop that was to become a Toronto landmark.
10. That he was still alive more than a century after his mother gave birth to him in a small, cold hut near Kiev had always impressed me too.
11. ¶As usual, over hot tea, Zayde and I chatted.
12. He asked me, as he always did, about my life—school, girlfriends, my parents, brother, sister.
13. These questions and answers served as a kind of prologue to the real discussion that would always follow.
14. I would ask him about his life in Russia and his role in the Russo-Japanese War of 1904-05.
15. Officially, he has been a drummer in the Russian Army; unofficially he taught fellow soldiers how to evade service by feigning all kinds of illnesses.
16. ¶Zayde would tell me how he escaped from Russia by a combination of bribery and good luck and came to Canada shortly afterwards.
17. “Russia no good, Canada wonderful,” he said.
18. He has theories about why he lived so long: “Never get excited, go for a walk.”
19. He told how it felt to be blind, as he had become in the last few years of his life.
20. “What can I do, I have lived a long life?”
21. He paused.
22. “But I would like to see the flowers and the birds, and I want to see the trees.”
23. ¶This time, however, my mission was not to be a grandson, but to act as his doctor and speak to him about his problem.
24. After an interminable period, I finally broached the subject, using techniques I had been learning in medical school.
25. ¶“Zayde,” I started uneasily, “you know, some old people sometimes get into trouble if they drink too much.”
26. “Oh, yes” he agreed.

27. "That's a real problem, you're right!"

28. Strike one.

29. ¶I tried again.

30. "An occasional drink doesn't hurt, but more than one a day is probably too much, don't you think?"

31. He smiled.

32. "Yes, absolutely!"

33. Strike two.

34. ¶"But Zayde," I blurted out, "we're afraid that you're going to fall down, break a leg, fracture your skull, have a heart attack!"

35. ¶"Hey, wait a minute, don't get excited.

36. Are you afraid that maybe I drink too much?" he asked.

37. ¶"Well, yes, sort of," I replied uncomfortably.

38. ¶"Now, now, don't worry about me" he said.

39. "I am okay.

40. *You* are the one that has to take care of yourself".

41. Zayde reached out for me, solicitously, his hands groping.

42. I felt like Jacob falsely seeking blessing from the dim-sighted Isaac.

43. "Watch your own health," Zayde continued.

44. "Be very careful!"

45. ¶"Why?" I asked, perplexed and a bit worried by his obvious concern for me.

46. ¶Looking out from his unseeing, yet far from expressionless, eyes—can a blind man's eyes twinkle?—he fixed me in a riveting gaze.

47. "Just remember," he said in such a low voice that I could barely hear him, "there are a lot more old drunks than old doctors."

48. ¶Today, nearly five years after his death, I look back on my visits with him with profound gratitude.

49. I reached adulthood while he still lived, and I crammed in as much talk with him as possible.

50. I knew the privilege could not last.

51. ¶My grandfather, who continued to drink what he wanted and never did become an alcoholic, was not an intellectual.

52. Although he could read and write Yiddish and spoke five languages, he could hardly spell an English word.

53. Most probably, he had never heard of Francis Bacon, yet the great English philosopher was certainly acquainted with the likes of my grandfather.

54. "The monuments of wit survive the monuments of power" was how Bacon would have summed him up.

4.3.2 The Lexical Structure

The following tables contain the lexical chains found in example (4-2):

Chain 1, Segment 1		
Word	Sentence	Lexical Chain
1. lucid	1	
2. alcohol	2	$(2, 1)^{5}_{0}$
3. indulge	3	$(3, 2)^{2}_{0}$ $(3, 1)^{T1}_{1}$
4. drink	3	$(4, 3)^{1}_{0}$ $(4, 2)^{1}_{0}$ $(4, 1)^{5}_{0}$
5. whiskey	3	$(5, 4)^{1}_{0}$ $(5, 1\text{-}2\text{-}3)^{T4}_{2}$
6. alcoholic	3	$(6, 2)^{0}_{0}$ $(6, 1)^{5}_{0}$ $(6, 3)^{2}_{0}$ $(6, 4)^{1}_{0}$ $(6, 5)^{T4}_{1}$
7. drinks	4	$(7, 4)^{0}_{0}$ $(7, 2\text{-}6)^{1}_{0}$ $(7, 3)^{2}_{0}$ $(7, 1\text{-}5)^{5}_{0}$
8. tippling	5	$(8, 2\text{-}3\text{-}4\text{-}6\text{-}7)^{1}_{0}$ $(8, 1\text{-}5)^{5}_{0}$

Chain 1, Segment 2		
Word	Sentence	Lexical Chain
9. drink	25	$(9, 4\text{-}7)^{0}_{0}$ $(9, 2\text{-}6\text{-}8)^{1}_{0}$ $(9, 3)^{2}_{0}$ $(9, 1\text{-}5)^{5}_{0}$

Chain 1, Segment 3		
Word	Sentence	Lexical Chain
10. drink	30	$(10, 4\text{-}7\text{-}9)^{0}_{0}$ $(10, 2\text{-}6\text{-}8)^{1}_{0}$ $(10, 3)^{2}_{0}$ $(10, 1\text{-}5)^{5}_{0}$

Chain 1, Segment 4		
Word	Sentence	Lexical Chain
11. drink	36	$(11, 4\text{-}7\text{-}9\text{-}10)^{0}_{0}$ $(11, 2\text{-}6\text{-}8)^{1}_{0}$ $(11, 3)^{2}_{0}$ $(11, 1\text{-}5)^{5}_{0}$

Chain 1, Segment 5		
Word	Sentence	Lexical Chain
12. drunks	47	$(12, 4\text{-}7\text{-}9\text{-}10\text{-}11)^{0}_{0}$ $(12, 2\text{-}6\text{-}8)^{1}_{0}$ $(12, 3)^{2}_{0}$ $(12, 1\text{-}5)^{5}_{0}$

Chain 1, Segment 6		
Word	Sentence	Lexical Chain
13. drink	51	$(13, 4\text{-}7\text{-}9\text{-}10\text{-}11\text{-}12)_0^0$ $(13, 2\text{-}6\text{-}8)_0^1$ $(13, 3)_0^2$ $(13, 1\text{-}5)_0^5$

Chain 1, Segment 7		
Word	Sentence	Lexical Chain
14. alcoholic	51	$(14, 2\text{-}6)_0^0$ $(14, 4\text{-}5\text{-}7\text{-}8\text{-}9\text{-}10\text{-}11\text{-}12\text{-}13)_0^1$ $(13, 3)_0^2$ $(13, 1\text{-}5)_0^5$

Chain 2, Segment 1		
Word	Sentence	Lexical Chain
1. maternal	1	
2. family	2	$(2, 1)_0^1$
3. aunt	3	$(3, 2)_0^1$ $(3, 1)_1^{T2}$
4. grandfather	3	$(4, 1\text{-}2)_0^1$ $(4, 3)_1^{T2}$
5. family	5	$(5, 1\text{-}2\text{-}3\text{-}4)_0^1$

Chain 2, Segment 2		
Word	Sentence	Lexical Chain
6. family	9	$(6, 1\text{-}2\text{-}3\text{-}4\text{-}5)_0^1$
7. mother	10	$(7, 1\text{-}2\text{-}4\text{-}5\text{-}6)_0^1$ $(7, 3)_1^{T2}$
8. parents	12	$(8, 1\text{-}2\text{-}4\text{-}5\text{-}6)_0^1$ $(8, 3)_1^{T2}$
9. brother	12	$(9, 2\text{-}3\text{-}5\text{-}6)_0^1$ $(9, 1\text{-}7\text{-}8)_1^{T2}$
10. sister	12	$(10, 2\text{-}3\text{-}5\text{-}6)_0^1$ $(10, 1\text{-}7\text{-}8)_1^{T2}$

Chain 2, Segment 3		
Word	Sentence	Lexical Chain
11. grandson	23	$(11, 5\text{-}6)_0^1$ $(11, 1\text{-}2\text{-}3\text{-}4\text{-}7\text{-}8\text{-}9\text{-}10)_1^{T2}$

Chain 2, Segment 4		
Word	Sentence	Lexical Chain
12. grandfather	51	$(12, 4)_0^0$ $(12, 2\text{-}5\text{-}6)_0^1$ $(12, 1\text{-}3\text{-}7\text{-}8\text{-}9\text{-}10\text{-}11)_1^{T2}$
13. grandfather	53	$(13, 4)_0^0$ $(13, 2\text{-}5\text{-}6)_0^1$ $(13, 1\text{-}3\text{-}7\text{-}8\text{-}9\text{-}10\text{-}11)_1^{T2}$

Chain 3		
Word	Sentence	Lexical Chain
1. house	6	
2. room	7	$(2, 1)_0^1$

Chain 4		
Word	Sentence	Lexical Chain
1. Canada	9	
2. Toronto	9	
3. Kiev	10	
4. Russia	14	
5. Russo-Japanese	14	
6. Russian	15	
7. Russia	16	
8. Canada	16	
9. Russia	17	
10. Canada	17	

Chain 5		
Word	Sentence	Lexical Chain
1. respect	8	
2. impressed	10	$(2, 1)_0^1$

Chain 6, Segment 1		
Word	Sentence	Lexical Chain
1. medical	5	
2. medical	5	$(2, 1)_0^0$

Chain 6, Segment 2		
Word	Sentence	Lexical Chain
3. doctor	23	$(3, 1\text{-}2)_0^1$
4. medical	24	$(4, 1\text{-}2)_0^0$ $(4, 3)_0^1$

Chain 6, Segment 3		
Word	Sentence	Lexical Chain
5. doctors	47	$(5, 3)_0^0$ $(5, 1\text{-}2\text{-}4)_0^1$

Chain 7		
Word	Sentence	Lexical Chain
1. asked	12	
2. questions	13	$(2, 1)_0^1$
3. answers	13	$(3, 1\text{-}2)_0^1$
4. discussion	13	$(4, 1\text{-}2)_0^1$ $(4, 3)_0^2$
5. ask	14	$(5, 1\text{-}2\text{-}3\text{-}4)_0^1$

Chain 8		
Word	Sentence	Lexical Chain
1. war	14	
2. drummer	15	$(2, 1)_0^1$
3. army	15	$(3, 1)_0^1$ $(3, 2)_1^{T1}$
4. soldiers	15	$(4, 1\text{-}2\text{-}3)_0^1$
5. service	15	$(5, 1\text{-}2\text{-}4)_0^1$ $(5, 3)_1^{T4}$

Chain 9		
Word	Sentence	Lexical Chain
1. evade	15	
2. feigning	15	$(2, 1)_0^2$
3. escaped	16	$(3, 1)_0^1$ $(3, 2)_1^{T1}$

Chain 10, Segment 1		
Word	Sentence	Lexical Chain
1. lived	1	
2. end	2	$(2, 1)^4_0$
3. died	2	$(3, 1)^4_0$ $(3, 2)^1_0$
4. lived	3	$(4, 1)^0_0$ $(4, 2\text{-}3)^4_0$

Chain 10, Segment 2		
Word	Sentence	Lexical Chain
5. alive	10	$(5, 1\text{-}4)^1_0$ $(5, 2\text{-}3)^4_0$
6. birth	10	$(6, 1\text{-}3\text{-}4\text{-}5)^{T2}_1$ $(6, 2)^4_0$
7. life	12	$(7, 1\text{-}4)^0_0$ $(7, 5)^1_0$ $(7, 2\text{-}3)^4_0$ $(7, 6)^{T2}_1$
8. life	14	$(8, 1\text{-}4\text{-}7)^0_0$ $(8, 5)^1_0$ $(8, 2\text{-}3)^4_0$ $(8, 6)^{T2}_1$
9. lived	18	$(9, 1\text{-}4\text{-}7\text{-}8)^0_0$ $(9, 5)^1_0$ $(9, 2\text{-}3)^4_0$ $(9, 6)^{T2}_1$
10. life	19	$(10, 1\text{-}4\text{-}7\text{-}8\text{-}9)^0_0$ $(10, 5)^1_0$ $(10, 2\text{-}3)^4_0$ $(10, 6)^{T2}_1$
11. lived	20	$(11, 1\text{-}4\text{-}7\text{-}8\text{-}9\text{-}10)^0_0$ $(11, 5)^1_0$ $(11, 2\text{-}3)^4_0$ $(11, 6)^{T2}_1$
12. life	20	$(12, 1\text{-}4\text{-}7\text{-}8\text{-}9\text{-}10\text{-}11)^0_0$ $(12, 5)^1_0$ $(12, 2\text{-}3)^4_0$ $(12, 6)^{T2}_1$

Chain 10, Segment 3		
Word	Sentence	Lexical Chain
13. death	48	$(13, 2\text{-}3)^1_0$ $(13, 1\text{-}4\text{-}5\text{-}7\text{-}8\text{-}9\text{-}10\text{-}11\text{-}12)^4_0$ $(13, 6)^{T2}_1$
14. lived	49	$(14, 1\text{-}4\text{-}7\text{-}8\text{-}9\text{-}10\text{-}11\text{-}12)^0_0$ $(14, 5)^1_0$ $(14, 2\text{-}3\text{-}13)^4_0$ $(14, 6)^{T2}_1$

Chain 11, Segment 1		
Word	Sentence	Lexical Chain
1. blind	19	
2. see	22	$(2, 1)^3_0$
3. see	22	$(3, 2)^0_0$ $(3, 1)^3_0$

Chain 11, Segment 2		
Word	Sentence	Lexical Chain
4. dim-sighted	42	
5. unseeing	46	$(5,\ 2\text{-}3)^0_0$ $(5,\ 1)^3_0$
6. eyes	46	$(6,\ 2\text{-}3\text{-}5)^1_0$ $(6,\ 1)^4_0$
7. twinkle	46	$(7,\ 2\text{-}3\text{-}5\text{-}6)^2_0$ $(7,\ 1)^3_0$
8. gaze	46	$(8,\ 2\text{-}3\text{-}5\text{-}6)^1_0$ $(8,\ 1)^4_0$ $(8,\ 7)^{T1}_1$

Chain 12, Segment 1		
Word	Sentence	Lexical Chain
1. uneasily	25	
2. trouble	25	$(2,\ 1)^1_0$
3. hurt	30	$(3,\ 1)^2_0$ $(3,\ 2)^1_0$
4. blurted	34	
5. afraid	34	$(5,\ 2\text{-}3)^2_0$ $(5,\ 1)^{T2}_1$
6. excited	35	$(6,\ 1\text{-}2\text{-}3\text{-}5)^2_0$
7. afraid	36	$(7,\ 5)^0_0$ $(7,\ 3)^1_0$ $(7,\ 2\text{-}6)^2_0$ $(7,\ 1)^{T2}_1$
8. uncomfortably	37	$(8,\ 1\text{-}2\text{-}3)^1_0$ $(8,\ 5\text{-}6\text{-}7)^2_0$
9. worry	38	$(9,\ 1\text{-}2\text{-}3\text{-}5\text{-}8)^1_0$ $(9,\ 6\text{-}7)^2_0$

Chain 12, Segment2		
Word	Sentence	Lexical Chain
10. perplexed	45	$(10,\ 2\text{-}3\text{-}8\text{-}9)^1_0$ $(10,\ 1\text{-}5\text{-}6\text{-}7)^2_0$
11. worried	45	$(11,\ 9)^0_0$ $(11,\ 1\text{-}2\text{-}3\text{-}8\text{-}10)^1_0$ $(11,\ 5\text{-}6\text{-}7)^2_0$
12. concern	45	$(12,\ 1\text{-}2\text{-}8\text{-}9\text{-}11)^1_0$ $(12,\ 3\text{-}5\text{-}6\text{-}7\text{-}10)^2_0$

Chain 13		
Word	Sentence	Lexical Chain
1. gratitude	48	
2. privilege	50	

Chain 14		
Word	Sentence	Lexical Chain
1. intellectual	51	
2. philosopher	53	2101
3. wit	54	$(3, 1\text{-}2)^1_0$
4. Francis Bacon	53	
5. Bacon	54	$(5, 4)^0_0$

Chain 15		
Word	Sentence	Lexical Chain
1. Yiddish	52	
2. languages	52	
3. English	52	$(2, 1)^1_0$
4. English	53	$(3, 2)^0_0$ $(3, 1)^1_0$

Chain 16		
Word	Sentence	Lexical Chain
1. strike one	28	
2. strike two	33	

Chain 17		
Word	Sentence	Lexical Chain
1. Zayde	2	
2. Zayde	5	
3. Zayde	8	
4. Zayde	11	
5. Zayde	16	
6. Zayde	25	
7. Zayde	34	
8. Zayde	41	
9. Zayde	44	

Chain 18		
Word	Sentence	Lexical Chain
1. assign	2	
2. dole	4	$(2, 1)_0^1$
3. delegated	5	$(3, 1)_0^1$ $(3, 2)_1^{T2}$

Chain 19		
Word	Sentence	Lexical Chain
1. climbing	7	
2. steep	7	$(2, 1)_0^1$
3. stairs	7	$(3, 1\text{-}2)_0^1$

4.3.3 The Intentional Structure

Figure 4.2 gives the intentional structure for this example.

Figure 4.2: The Intentional Structure of Example (4-2)

1 (1–54)
Describe the mission to confront the grandfather with his family perceived problems with alcohol, and in so doing, to describe the grandfather's long and useful life.

1.1 (1–5)
Discuss the family approach to the grandfather's perceived alcohol problem.

1.1.1 (1–3)
Describe the family concerns over the grandfather's drinking.

1.1.2 (4–5)
Describe how author was chosen by the family to confront the grandfather since he is a doctor.

1.2 (6–47)
Describe author's visit to his grandfather to tell him about his problem, and to extoll the virtues of his grandfather in the telling of this event.

1.2.1 (6–22)
Describe author's unhappiness with having to confront his grandfather with his problem with alcohol.

1.2.1.1 (6–7)

Describe author's confusion and unease with the imminent task of confronting his grandfather with his drinking problem.

1.2.1.2 (8–22)

Describe how much author repects his grandfather and how this makes him feel uncomfortable with the task of confronting his grandfather with an alcohol problem.

1.2.1.2.1 (8–10)

Describe the grandfather's marvelous work achievements.

1.2.1.2.2 (11–16)

Describe the grandfather's commendable activity during the war.

1.2.1.2.3 (17–22)

Describe the wonderful length of years author's grandfather has lived.

1.2.2 (23–47)

Describe the discussion with author's grandfather about his family perceived alcohol problem.

1.2.2.1 (23–33)

Describe author's first two failed attempts at starting the discussion of the alcohol problem.

1.2.2.2 (34–37)

Describe how author finally blurted out the problem.

1.2.2.3 (38–47)

Describe that the grandfather's reaction was concern for the author (there are more old drunks that old doctors).

1.3 (48–54)

Describe author's fond remembrances of his grandfather.

1.3.1 (48–50)

Describe author's gratitude for the privilege of knowing his grandfather.

1.3.2 (51)

Describe how the grandfather never became an alcoholic.

1.3.3 (52–54)

Describe what a great man the grandfather was even though he was not an intellectual scholar.

4.3.4 The Correspondences between Lexical and Intentional Structures

The following table gives the correspondences between the lexical chains and intentions of example (4-2):

Chain	Chain Range	Intention	Intention Range
1.1	1–5	1.1	1–5
1	1–51	1	1–54
2.1	1–5	1.1	1–5
2.2	9–12	1.2.1.2.1	8–10
2.3	23	start of 1.2.2	23
2.4	51–53	1.3	48–54
3	6–7	1.2.1.1	6–7
4	9–17	1.2.1.2	8–22
5	8–10	1.2.1.2.1	8–10
6.1	5	1.1.2	4–5
6.2	23–24	start of 1.2.2	23
6.3	47	end of 1.2	47
7,8,9	12–16	1.2.1.2.2	11–16
10.1	1–3	1.1.1	1–3
10.2	10–20	1.2.1.2	8–22
10.3	48–49	1.3.1	48–50
11.1	19–22	1.2.1.1.3	17–22
11.2	42–46	1.2.2.3	38–47
12.1	25–38	1.2.2	23–38
12.2	45	end of 1.2.2	47
13	48–50	1.3.1	48–50
14	51–54	1.3.3	52–54
15	52–53	1.3.3	52–54
16	28–33	1.2.2.1	23–33
17	1–44	1	1–54
18	2–5	1.1	1–5
19	7	1.2.1.1	6–7

Chain 1.1 corresponds to intention 1.1 exactly, meaning that they both start and end on the same sentence boundary. Chain 1, however, continues throughout almost all of the example, mostly in the form of one-word chain returns. This is taken to indicate unity throughout the entire example. Each of these returns except segment 3 corresponds closely to an intentional boundary. Segments 5 and 6 do exactly, segment 4 is one sentence from a boundary, and segment 2 is two sentences from a boundary.

Chain 2 is another long chain that is made up of several returns. Taken as a whole, the chain segments indicate unity throughout the entire example. Consideration of each individual segment also provides clues to the intentional structure. Chain 2.1 corresponds exactly to intention 1.1. Chain 2.2 corresponds to intention 1.2.1.2.1. Notice that the lexical chain 2.2 actually overlaps into the sentences of intention 1.2.1.2.2. Chain 2.2 runs to sentence 12, and intention 1.2.1.2.2 starts at sentence 11. Sentences 11 and 12 of intention 1.2.1.2.2 could possibly warrant a separate intention, since they are not directly about either Zayde's life or his role in the war. It seems that fuzziness in determining intentional boundaries is reflected by lexical chain boundaries overlapping intentional boundaries. Chain 2.3 corresponds with the start of intention 1.2.2. This is not a coincidence but a lexical indication that the diversion to the grandfather's life-story is over, and the grandson is now going to tell of his confrontation with his grandfather. Similarly, chain 2.4 indicates the end of intention 1.2—the discussion with Zayde and the grandson of the drinking problem, and the start of intention 1.3—the grandson expounding on his grandfather's virtures.

Chain 3 corresponds exactly to intention 1.2.1.1. Chain 4 corresponds to intention 1.2.1.2, although it stops short of the end of the intention. Chain 5 corresponds to intention 1.2.1.2.1.

Chain 6 consists of three segments that span sentences 5 to 47, and so provides a clue to the unity of the intention 1.2, which runs from sentences 6 to 47. Chain segment 6.1 corresponds to intention 1.1.2. Chain segment 6.2 corresponds to the start of intention 1.2.2, and segment 6.3 corresponds to the end of intention 1.2. Hence the chain returns in chain 6 unify a high-level intention that consists of several lower-level intentions. They also occur at intentional boundaries.

Chains 7, 8, and 9 are grouped together to collectively correspond to one intention—intention 1.2.1.2.2. The chains are short in terms of the number of sentences they span, and they overlap with each other. Both of these facts suggest that the chains should be grouped together as indicative of a structural unit of text. This example illustrates how and why this tool must be integrated with other text analysis tools. It is necessary in this case to validate the correspondence. Not surprisingly, the three chains are strongly related in the meaning context of the text. Chain 7 contains {*asked, questions*}, chain 8 contains {*war, army*} and chain 9 contains {*evade, escaped*}. Chains 8 and 9 are specifically what the *questions* were *asked* about. Also, *evade* and *escape* from chain 9 are in the context of *war* from chain 8.

Chain 10.1 corresponds to intention 1.1.1, chain 10.2 corresponds to intention 1.2.1.2, and chain 10.3 corresponds exactly to intention 1.3.1. Chain 10 as a whole runs through the entire text from sentences 1 to 49. There is no intention spanning sentences 1 to 49. In this case the chain seems to give unity to the entire example, even though it does not quite span all of it. This again shows that the lexical chain information provided by this tool must be integrated with other sources of textual information to provide a complete structural analysis of the text. The returns in chain 10 both occur at intentional boundaries.

Chain 11.1 corresponds to intention 1.2.1.2.3. Chain 11.2 corresponds to intention 1.2.2.3. Unfortunately, chain segment 11.2 is a counter-intuitive return to chain 11. The analysis resulting from considering 11.2 as a return would be that the entire span of sentences from intention 1.2.1.2.3 to intention 1.2.2.3 should be linked together as a structural unit, and in fact they are both a part of intention 1.2. However, this was not intuitively apparent, nor is the tie very clean, since intention 1.2.1.2.3 is deeply embedded near the end of intention 1.2.1, and intention 1.2.2.3 is at the end of intention 1.2.

Chain 12.1 corresponds to intentions 1.2.2.1 and 1.2.2.2. The return segment, 12.2, corresponds to the end of intention 1.2.2.3. Also, this return serves to tie together all of intention 1.2.2 which spans sentences 23 to 47. Chain 12 spans sentences 25 to 45.

Chain 13 corresponds exactly to intention 1.3.1. Chain 14 corresponds to intention 1.3.3. Chain 15 is entirely contained in chain 14, and with a finer granularity of intentional structure analysis, could correspond to a nested intention. This nested intention could be specifically about a language aspect of intellect.

Chain 16 corresponds to intention 1.2.2.1. Even though the start of the chain is five (short) sentences after the start of the intention, the existence of a unique chain suggests a unique intention in that textual area. Chain 16 and intention 1.2.2.1 both end at sentence 33.

Chain 17 (containing only the word *Zayde*) corresponds to intention 1. With the addition of lexical rendering (see section 3.3) there would be no breaks in the chain, as pronouns referring to *Zayde* are used extensively throughout the example. The chain returns in this case do not correspond to intentional boundaries. The existence of so many short returns spaced throughout the entire example seems only to tie the entire text together.

Chain 18 corresponds to intention 1.1. Chain 19, consisting of three words in one sentence, would seem to suggest an intentional boundary, and in fact is within one sentence of a major boundary between intentions 1.1 and 1.2.

A major conclusion to be drawn from this example is that in many cases the lexical chain boundaries are within one or two sentences of the corresponding intentions. Sometimes the lexical chains are shorter than their corresponding intention, but they are still useful in suggesting the presence of a unique structural unit in their textual vicinity.

Chain returns in this example are again strong indicators of both structural boundaries and structural ties linking low-level intentions together into a higher-level one.

In one case, the combination of short overlapping chains (chains 7, 8, and 9) corresponded to one intention. In two cases short chains (chains 3 and 17) provided an indication of a structural boundary.

4.4 Example (4-3)

4.4.1 The Text

Here is the text of example (4-3), the first section of an article in the *New Yorker* magazine, July 27, 1987, by Arlene Croce, entitled "The Bolshoi bows in": [4]

(4-3)

1. ¶New Yorkers who went to the Bolshoi Ballet this summer at the Metropolitan Opera House were put through an elaborate security net.
2. Standing in long lines, they submitted to having their handbags searched and their persons surveyed by metal detectors, then to being penned inside the Met

[4] Reprinted by permission; © 1987 Arlene Croce. Originally in *The New Yorker*

for the entire performance.

3. No going out-of-doors into Lincoln Center Plaza during intermissions, not for an ice-cream cone, not even to smoke.
4. When the performance started again, every ticket stub was rechecked, in case someone had got into the building without a ticket.
5. Since the Met has four thousand and sixty-five seats and every one of them had been sold (at a top of sixty dollars), the security procedures were a great bother, especially during the congested intermissions, but they were preferable to the type of assault on the audience which had taken place on the first night of the Moiseyev Dance Company's season at the same theatre last fall, when members of the Jewish Defense League released tear gas inside the house.
6. On the first night of the Bolshoi season, a bomb scare caused the curtain to be delayed for an hour.
7. The security checks that went on thereafter added considerably to the running time of every performance, but no one seemed to mind getting out late.
8. Many people even stayed on to applaud, whipped up as much, no doubt, by the tension of the occasion as by the performance.
9. ¶And by the Bolshoi mystique.
10. This mystique, the joint creation of balletomanes, P.R. men, and politicians, has at its root a fixed belief in the supremacy of Russian ballet and the benefits of cultural exchange.
11. Applauding Soviet dancers—letting them know we approve of and understand their art—is supposed to be good not only for one's own cultural health but for the health of nations.
12. The Bolshoi mystique has always been a factor in the company's success here.
13. People gripped by it don't just applaud dancing; they applaud demonstratively, caringly, as if casting a vote.
14. They *come* to applaud, and this being the first Bolshoi visit in eight years, they sounded off with an enthusiasm that would not be dampened—not by bomb scares and security checks and boosted ticket prices, not by a heat wave that blanketed the city, not even by the ballets of Yuri Grigorovich.
15. Every night, they rushed to the opera house, passed their security test and cheered the cause of art and peace.
16. ¶So eager was the response of these good citizens that it became disruptive, constantly anticipating cues for applause.
17. And the Bolshoi dancers are very particular about cues.
18. When the applause wasn't forthcoming at the right moment, the fact that it had already been granted made no difference; they demanded it anyway.
19. When it died, they walked out and revived it by bowing some more.
20. They bowed between the acts of a ballet, taking half-risen audiences by surprise, and they bowed during the coda of a classical grand pas de deux, rudely cutting off the music to do so.

21. Plainly, the Bolshoi is used to an applause *routine*.

22. Dancers with so little audience rapport that they can't adjust a simple curtain call are very strange to see, but the failing is symbolic of the insecurity and the lack of finesse that have afflicted the Bolshoi command all these years.

23. (I don't mean to exempt the Kirov, in whose dancers the same bad habits are ingrained; but the Kirov has consistently held itself more aloof from the public than the Bolshoi.)

4.4.2 The Lexical Structure

The following tables show the lexical chains found in example (4-3):

Chain 1		
Word	Sentence	Lexical Chain
1. New Yorkers	1	
2. Metropolitan Opera House	1	
3. Met	2	
4. Lincoln Centre Plaza	3	
5. Met	5	$(5, 3)^0_0$

Chain 2, Segment 1		
Word	Sentence	Lexical Chain
1. security	1	
2. submitted	2	$(2, 1)^{5}_{0}$
3. searched	2	
4. surveyed	2	$(4, 1)^{2}_{0}$ $(4, 2)^{T1}_{1}$
5. metal detectors	2	
6. penned	2	$(6, 1)^{1}_{0}$ $(6, 2)^{2}_{0}$ $(6, 4)^{T1}_{1}$
7. rechecked	4	$(7, 1\text{-}6)^{1}_{0}$ $(7, 2)^{2}_{0}$ $(7, 4)^{T1}_{1}$
8. security	5	$(8, 1)^{0}_{0}$ $(8, 6\text{-}7)^{1}_{0}$ $(8, 4)^{2}_{0}$ $(8, 2)^{5}_{0}$
9. bother	5	$(9, 3)^{1}_{0}$ $(9, 4)^{1}_{0}$ $(9, 2)^{T2}_{1}$ $(9, 1\text{-}8)^{T4}_{1}$
10. congested	5	
11. assault	5	$(11, 1)^{2}_{0}$ $(11, 2\text{-}4\text{-}6\text{-}7\text{-}8)^{T1}_{1}$
12. tear gas	5	$(12, 3\text{-}9)^{1}_{0}$ $(12, 4)^{T9}_{1}$
13. bomb scare	6	$(13, 1)^{1}_{0}$ $(13, 4\text{-}6\text{-}7)^{2}_{0}$ $(13, 2)^{T6}_{1}$ $(13, 8)^{T6}_{1}$ $(13, 9)^{T4}_{1}$ $(13, 11)^{T1}_{1}$
14. delayed	6	$(14, 7)^{2}_{0}$ $(14, 1\text{-}2\text{-}6\text{-}8\text{-}13)^{T7}_{1}$
15. security	7	$(15, 1\text{-}8)^{0}_{0}$ $(15, 6\text{-}7)^{1}_{0}$ $(15, 4)^{2}_{0}$ $(15, 2)^{5}_{0}$ $(15, 11\text{-}13)^{T1}_{1}$ $(15, 9)^{T4}_{1}$
16. tension	8	

Chain 2, Segment 2

Word	Sentence	Lexical Chain
17. bomb scares	14	$(17, 13)_0^0$ $(17, 1)_0^1$ $(17, 4\text{-}6\text{-}7)_0^2$ $(17, 2)_1^{T6}$ $(17, 8)_1^{T6}$ $(17, 9)_1^{T4}$ $(17, 11)_1^{T1}$ $(17, 14)_1^{T7}$ $(17, 15)_1^{T1}$
18. security	14	$(18, 1\text{-}8\text{-}15)_0^0$ $(18, 6\text{-}7)_0^1$ $(18, 2)_0^5$ $(18, 4)_0^2$ $(18, 9)_1^{T4}$ $(18, 11\text{-}13\text{-}17)_1^{T1}$
19. security	15	$(19, 1\text{-}8\text{-}15\text{-}18)_0^0$ $(19, 6\text{-}7)_0^1$ $(19, 2)_0^5$ $(19, 4)_0^2$ $(19, 9)_1^{T4}$ $(19, 11\text{-}13\text{-}17)_1^{T1}$
20. disruptive	16	

Chain 2, Segment 3

Word	Sentence	Lexical Chain
21. rudely	20	$(21, 3\text{-}9\text{-}12)_0^2$ $(21, 4)_1^{T9}$
22. strange	22	$(22, 20)_0^1$
23. failing	22	
24. insecurity	22	$(24, 1\text{-}8\text{-}15\text{-}18\text{-}19)_0^0$ $(24, 6\text{-}7)_0^1$ $(24, 2)_0^5$ $(24, 4)_0^2$ $(24, 9)_1^{T4}$ $(24, 11\text{-}13\text{-}17)_1^{T1}$
25. lack	22	$(25, 22)_0^1$
26. afflicted	22	$(26, 3\text{-}9)_0^1$ $(26, 12\text{-}21)_0^2$ $(26, 4)_1^{T9}$ $(26, 21)_1^{T3}$
27. bad	23	$(27, 21)_0^1$ $(27, 3\text{-}9\text{-}12\text{-}26)_1^{T2}$

Chain 3		
Word	Sentence	Lexical Chain
1. Bolshoi Ballet	1	
2. performance	2	$(2, 1)^1_0$
3. intermissions	3	
4. performance	4	$(4, 2)^0_0$ $(4, 1)^1_0$
5. ticket	4	
6. ticket	4	$(6, 5)^0_0$
7. intermissions	5	
8. audience	5	
9. dance	5	$(9, 1)^1_0$ $(9, 2\text{-}4)^{T1}_1$
10. theatre	5	$(10, 1\text{-}2\text{-}9)^1_0$ $(10, 4)^{T1}_1$
11. Bolshoi	6	$(11, 1)^0_0$
12. season	6	
13. curtain	6	$(13, 1\text{-}2\text{-}4\text{-}10)^1_0$ $(13, 9)^{T10}_1$
14. performance	7	$(14, 2\text{-}4)^0_0$ $(14, 1)^1_0$ $(14, 9\text{-}10\text{-}13)^{T1}_1$
15. applaud	8	
16. performance	8	$(16, 2\text{--}4\text{-}14)^0_0$ $(16, 1)^1_0$ $(16, 9\text{-}10\text{-}13)^{T1}_1$
17. Bolshoi	9	$(17, 1\text{-}11)^0_0$
18. balletomanes	10	$(18, 1)^0_0$ $(18, 2\text{-}4\text{-}9\text{-}10\text{-}13\text{-}14\text{-}16)^1_0$
19. ballet	10	$(19, 1\text{-}18)^0_0$ $(19, 2\text{-}4\text{-}9\text{-}10\text{-}13\text{-}14\text{-}16)^1_0$
20. cultural	10	$(20, 9)^2_0$ $(20, 1\text{-}10\text{-}18\text{-}19)^{T9}_1$

Chain 3 (continued)		
Word	Sentence	Lexical Chain
21. applauding	11	$(21, 15)_0^0$
22. dancers	11	$(22, 9)_0^0$ $(22, 1\text{-}10)_0^1$ $(22, 20)_0^2$ $(22, 2\text{-}4\text{-}14\text{-}16\text{-}18\text{-}19)_1^{T10}$ $(22, 13)_1^{T1}$
23. art	11	$(23, 20)_0^1$ $(23, 12)_0^2$ $(23, 9\text{-}22)_1^{T20}$
24. cultural	24	$(24, 20)_0^0$ $(24, 23)_0^1$ $(24, 9\text{-}12)_0^2$ $(24, 10\text{-}18\text{-}19)_1^{T9}$
25. Bolshoi	12	$(25, 1\text{-}11\text{-}17)_0^0$
26. company	12	$(26, 1\text{-}2\text{-}4\text{-}9\text{-}10\text{-}13\text{-}14\text{-}16\text{-}22)_0^1$ $(26, 8\text{-}20\text{-}24)_0^2$ $(26, 18\text{-}19)_1^{T2}$
27. applaud	13	$(27, 15\text{-}21)_0^0$
28. dancing	13	$(28, 9\text{-}22)_0^0$ $(28, 1\text{-}10)_0^1$ $(28, 20)_0^2$ $(28, 2\text{-}4\text{-}14\text{-}16\text{-}18\text{-}19)_1^{T10}$ $(28, 13\text{-}26)_1^{T1}$ $(28, 23)_1^{T20}$ $(28, 24)_1^{T9}$
29. applaud	13	$(29, 15\text{-}21\text{-}27)_0^0$
30. applaud	14	$(30, 15\text{-}21\text{-}27\text{-}29)_0^0$
31. Bolshoi	14	$(31, 1\text{-}11\text{-}17\text{-}25)_0^0$
32. enthusiasm	14	$(32, 15\text{-}27\text{-}29\text{-}30)_0^2$
33. ticket	14	$(33, 5\text{-}6)_0^0$
34. ballets	14	$(34, 1\text{-}18\text{-}19)_0^0$ $(34, 2\text{-}4\text{-}9\text{-}10\text{-}13\text{-}14\text{-}16\text{-}22\text{-}26)_0^1$ $(34, 20)_1^{T28}$
35. opera	15	$(35, 1\text{-}2\text{-}4\text{-}10\text{-}13\text{-}14\text{-}16\text{-}18\text{-}26\text{-}34)_0^1$ $(35, 28)_1^{T1}$ $(35, 9\text{-}19\text{-}20\text{-}22)_1^{T28}$
36. cheered	15	$(36, 1\text{-}2\text{-}4\text{-}9\text{-}10\text{-}14\text{-}18\text{-}20\text{-}22\text{-}24\text{-}28\text{-}34)_0^1$ $(36, 13\text{-}16\text{-}19)_1^{T34}$
37. art	15	$(37, 23)_0^0$ $(37, 20)_0^1$ $(37, 12\text{-}24)_0^2$ $(37, 9\text{-}22)_1^{T20}$
38. eager	16	$(38, 32)_0^1$ $(38, 15\text{-}27\text{-}29\text{-}30)_1^{T32}$
39. cues	16	$(39, 2\text{-}4\text{-}14\text{-}16\text{-}23\text{-}37)_0^1$ $(39, 12\text{-}20\text{-}24)_1^{T37}$

Chain 3 (continued)		
Word	Sentence	Lexical Chain
40. applause	16	(40, 15-21-27-29-30)$^{0}_{0}$ (40, 36)$^{1}_{0}$ (40, 2-4-9-14-18-20-22-24-28-34)$^{T36}_{1}$
41. Bolshoi	17	(41, 1-11-17-25-31)$^{0}_{0}$
42. dancers	17	(42, 9-22-28)$^{0}_{0}$ (42, 1-10-34)$^{1}_{0}$ (42, 20)$^{2}_{0}$ (42, 2-4-14-16-18-19)$^{T10}_{1}$ (42, 13-26)$^{T1}_{1}$ (42, 23)$^{T20}_{1}$ (42, 24)$^{T9}_{1}$
43. cues	17	(43, 39)$^{0}_{0}$ (43, 2-4-14-16-23-37)$^{1}_{0}$ (43, 9-12-20-22-24)$^{T37}_{1}$
44. applause	18	(44, 15-21-27-29-30-40)$^{0}_{0}$ (44, 36)$^{1}_{0}$ (44, 2-4-9-14-18-20-22-24-28-34)$^{T36}_{1}$
45. bowing	19	(45, 9-22-24-28-42)$^{2}_{0}$ (45, 12-23)$^{T24}_{1}$ (45, 1-10-20)$^{T42}_{1}$
46. bowed	20	(46, 9-22-24-28-42)$^{2}_{0}$ (46, 12-23)$^{T24}_{1}$ (46, 1-10-20)$^{T42}_{1}$ (46, 45)$^{0}_{0}$
47. act	20	(47, 1-2-4-10-12-13-14-16-19-20-23-26-34-37-39-43)$^{1}_{0}$ (47, 9-18-22-28)$^{T34}_{1}$ (47, 42)$^{T1}_{1}$
48. ballet	20	(48, 1-18-19-34)$^{0}_{0}$ (48, 2-4-9-10-14-16-22-26-28-35-36-42-47)$^{1}_{0}$ (48, 12-13-23-37-39-43)$^{T47}_{1}$ (48, 24-45-46)$^{T42}_{1}$
49. audiences	20	(49, 8)$^{0}_{0}$ (49, 36)$^{1}_{0}$ (49, 9-22-24-28-42)$^{2}_{0}$ (49, 1-2-4-9-14-16-18-22-34)$^{T36}_{1}$ (49, 1-10-20-45-46-47-48)$^{T42}_{1}$
50. surprise	20	
51. bowed	20	(51, 45-46)$^{0}_{0}$ (51, 9-22-24-28-42)$^{2}_{0}$ (51, 12-23)$^{T24}_{1}$ (51, 1-10-20)$^{T42}_{1}$
52. pas de deux	20	
53. Bolshoi	21	(53, 1-11-17-25-31)$^{0}_{0}$
54. applause	21	(54, 15-21-27-29-30-40-44)$^{0}_{0}$ (54, 36)$^{1}_{0}$ (54, 2-4-9-14-18-20-22-24-28-34-48-49)$^{T36}_{1}$
55. dancers	22	(55, 9-22-28-42)$^{0}_{0}$ (55, 1-10-34)$^{1}_{0}$ (55, 20)$^{2}_{0}$ (55, 2-4-14-16-18-19)$^{T10}_{1}$ (55, 13-26)$^{T1}_{1}$ (55, 23)$^{T20}_{1}$ (55, 24)$^{T9}_{1}$ (55, 51)$^{2}_{0}$ (55, 45-46)$^{T51}_{1}$

Chain 3 (continued)		
Word	Sentence	Lexical Chain
56. audience	22	$(56, 8-20)^{0}_{0}$ $(56, 36)^{1}_{0}$ $(56, 9-22-24-28-42-49)^{2}_{0}$ $(56, 1-2-4-9-14-16-18-22-34)^{T36}_{1}$ $(56, 1-10-20-45-46-47-48)^{T42}_{1}$ $(56, 51)^{2}_{0}$ $(56, 55)^{1}_{0}$
57. curtain	22	$(57, 13)^{0}_{0}$ $(57, 1-2-4-10-14-16-35-4748)^{1}_{0}$ $(57, 12-19-20-23-26-34-37-39-43)^{T47}_{1}$ $(57, 9)^{T10}_{1}$ $(57, 18-36)^{T48}_{1}$
58. Bolshoi	22	$(58, 1-11--17-25-31-53)^{0}_{0}$
59. Kirov	23	
60. dancers	23	$(60, 9-22-28-42-55)^{0}_{0}$ $(60, 1-10-34-57)^{1}_{0}$ $(60, 20)^{2}_{0}$ $(60, 2-4-14-16-18-19)^{T10}_{1}$ $(60, 13-26)^{T1}_{1}$ $(60, 23)^{T20}_{1}$ $(60, 24)^{T9}_{1}$ $(60, 51)^{2}_{0}$ $(60, 45-46)^{T51}_{1}$
61. Kirov	23	$(61, 59)^{0}_{0}$
62. Bolshoi	23	$(62, 1-11-17-25-31-53-58)^{0}_{0}$

Chain 4		
Word	Sentence	Lexical Chain
1. mystique	9	
2. mystique	10	$(2, 1)^0_0$
3. mystique	12	$(3, 1\text{-}2)^0_0$

Chain 5		
Word	Sentence	Lexical Chain
1. rushed	15	
2. eager	16	$(2, 1)^1_0$
3. anticipating	16	$(3, 1)^2_0$ $(3, 2)^1_0$

Chain 6		
Word	Sentence	Lexical Chain
1. health	11	
2. health	11	$(2, 1)^0_0$

4.4.3 The Intentional Structure

Figure 4.3 gives the intentional structure for this example.

Figure 4.2: The Intentional Structure of Example (4-3)

- 1 (1–23)
  Describe how extreme security measures do not dampen the fervour with which ballet audiences in New York City applaud the Bolshoi, and how this is what the Bolshoi expects and in in fact demands.
  - 1.1 (1–15)
    Describe the effects of extreme security measures at the Metropolitan Opera House in New York City.
    - 1.1.1 (1–8)
      Describe what the security measures were.
    - 1.1.2 (9–15)
      Describe the audience enthusiasm due to the mystique and image of supremacy of the Bolshoi Ballet, and how security measures do not dampen this
      - 1.1.2.1 (9–13)
        Describe the enthusiasm to applaud due to the mystique and supremacy image of the Bolshoi.
      - 1.1.2.2 (14–15)
        Describe how the enthusiasm is not dampened by security.
  - 1.2 (16–23)
    Describe the lack of rapport between the audience and the dancers.
    - 1.2.1 (16–21)
      Describe the disruptive audience applause coupled with the irritating and too frequent cues given by the Bolshoi.
    - 1.2.2 (22–23)
      Describe how the disruptive applause indicates a failure of the Bolshoi and a lack of finesse in the company.

4.4.4 The Correspondences between Lexical and Intentional Structures

The following table gives the correspondences between the lexical chains and intentions of example (4-3):

Chain	Chain Range	Intention	Intention Range
1	1–5	1.1.1	1–8
2.1	1–8	1.1.1	1–8
2.2	14–16	1.1.2.2	14–5
2.3	20–23	1.2.2	22–23
3	1–23	1	1–23
4	9–13	1.1.2.1	9–13
5	15–16	1.1.2.2	14–15
6	11	1.1.2.1	9–13

Chain 1 corresponds to intention 1.1.1, although it is three sentences short of spanning the entire intention.

Chain 2.1 corresponds exactly to intention 1.1.1. Chain 2.2 corresponds to intention 1.1.2.2. It is also a return to chain 2, and so indicates a unity in the text extending from intention 1.1.1 to intention 1.1.2.2, hence pulling together intention 1.1. Chain 2.3 corresponds to intention 1.2. It is not considered to be a return to chain 2 (also see section 3.3.2 for an explanation) since it is not intuitively or semantically linked to the rest of chain 2. This situation is one in which the use of other textual information such as semantics must be integrated with the lexical chain information. Specifically, *disruptive* in sentence 16 is lexically related to chain segments 2.1 and 2.2, however, *disruptive* is used to describe the audience applause, and not the security measures. The use of this information would enable segment 2.3 to not be considered a return to chain 2.

Chain 3 runs steadily throughout the entire example, and so corresponds to intention 1. Within chain 3, variations of the word *applause* are repeated seven times between sentences 9 and 18, indicating the unity between intentions 1.1.2.1 and 1.1.2.2. *Bowed* is repeated three times in sentences 19 and 20, suggesting the presence of a separate structural unit of text in that vicinity. Specifically, it indicates the presence of intention 1.2.1. This analysis of repeated words is a case of chain strength information (repetition increases strength) supplementing the use of lexical chain information.

Chain 4 corresponds to intention 1.1.2.1 exactly. Chain 5 corresponds to intention 1.1.2.2. Chain 6 consists of a two-word reiteration in one sentence that does not seem to correspond clearly to any intention except perhaps 1.1.2.1. At this rather high granularity of intentional analysis, it is not surprising that a chain spanning one sentence does not have a more closely corresponding intention.

To conclude, this example again shows that the lexical chains are good indicators of intentional structure. Even though they do not always correspond exactly to the intentions in terms of specific sentences spanned, they are usually close (one to two sentences off). The example contains a chain return analysis (chain 2) showing that the return indicates both structural unity between low level intentions, and intentional boundaries. A case of chain strength indicating structural unity appears in chain 3.

4.5 Example (4-4)

4.5.1 The Text

Here is the text of example (4-4), the first part of an article in the *Life* section of the *Toronto Star*, December 31, 1987, entitled "Is Raisa a realistic Soviet role model?":[5]

(4-4)

1. ¶"Not since the czars ruled Russia has a woman maintained such a high profile in the Soviet society ...."
2. ¶This lead-in sentence of a popular talk-show host echoed throughout North America in a variety of permutations during the recent Gorbachev-Reagan summit.
3. ¶The slim, educated and outspoken Raisa Gorbachev has captured the attention of America, stirring considerable controversy by her reported rudeness to Nancy Reagan and her visible—and largely misunderstood—role in the public relations package the Soviets brought to the summit.
4. ¶To many people here, she represented the "new type of Soviet woman," an equal partner with her husband in affairs of the state, a professional woman who at the same time is appreciative of the finer things in life, including fur coats and American Express.
5. In short, a role model for Soviet women, as the talk-show host put it.
6. It is true that there has not been such a high-profile Soviet woman in many years.
7. But this carefully cultivated image of the "new Soviet woman" is strictly for export, an item manufactured for Western consumption.
8. ¶To understand why Mrs.'Gorvachev is not a suitable role model for Soviet women, one need only look at the reality of daily life for women in the U.S.S.R., and their position in a conservative, male-dominated society where they carry the overwhelming share of hardships.
9. They do have equal rights under Soviet law, a fact Soviet propagandists proudly point out as a proof of superiority of the socialist way of life.
10. Law, however, often has little to do with reality in the U.S.S.R.
11. And the reality—according to the official magazine of the Communist Party, Kommunist—is that in industry women are often confined to menial jobs, for example, lifting objects up to 54 kilograms (120 pounds), moving perhaps 10 tonnes during one shift.

[5] © The Washington Post. Reprinted by permission.

12. ¶The situation is not going to improve soon because the real power—the power to make policy decisions—remains an exclusively male preserve.

13. For instance, women make up 40 per cent of the academic professions, but there are fewer than 2 per cent of them among members of the Soviet Academy of Science, the power centre of the huge academic bureaucracy.

14. ¶And when it comes to real decision-making power, women have none.

15. Only once in Soviet history has there been a woman on the Politburo.

16. Her tenure was brief: She was brought into this inner sanctum of power by Nikita Khrushchev as part of his push for social reform and was removed swiftly as an unwelcome intruder once he fell from favor.

17. The odd twist is this: Soviet women cannot even fight for greater political clout because theoretically they already have all the rights granted to them by the Soviet constitution!

18. ¶So although there may be no legal roadblocks to prevent a woman from rising up the economic and social ladder, most simply give up.

19. Ninety per cent of Soviet women work full-time.

20. Housework is a second full-time job, almost never shared by the husband.

21. ¶It is true that in the United States, too, the salary of a working wife and mother can make a difference and is not, strictly speaking, "optional."

22. ¶Most women here don't work to be "liberated," but to make money and help support the family.

23. ¶But their salaries often make possible a higher standard of living, not mere survival.

24. In the Soviet Union, many women would like to be able to stay home and raise a family, but the reality of life leaves them with no choice.

25. ¶The ability, therefore, to have a job—as Mrs. Gorbachev did—that does not require being there full-time becomes a status symbol, generally bespeaking the power status of the husband.

26. Many academic positions allow for a day or two a week of "sabbatical work"—a euphemism for "staying home."

27. ¶Graduate study is the best time of all.

28. There are three years on the state stipend that is close to an average worker's wage.

29. And there are no courses or exams to take, except qualifying pre-enrolment exams, so all that remains to do is to write a dissertation and two or three publishable papers.

30. From my own experience of doing such study I know that, at least in social sciences, all this can be accomplished within six months.

31. ¶Therefore, foreign languages, publishing, social sciences and college-level teaching are favorites among wives of the power elite—and among their husbands who have enough clout to make sure that their wives' job demands to not take too much time away from their wifely duties.

4.5.2 The Lexical Structure

The following tables show the lexical chains found in example (4-4):

Chain 1, Segment 1		
Word	Sentence	Lexical Chain
1. czars	1	
2. ruled	1	$(2, 1)^2_0$
3. Russia	1	
4. Soviet	1	
5. North America	2	
6. Gorbachev-Reagan	2	
7. Gorbachev	3	$(7, 6)^0_0$
8. America	3	$(8, 5)^0_0$
9. Reagan	3	$(9, 6)^0_0$
10. Soviets	3	$(10, 4)^0_0$
11. Soviet	4	$(11, 10\text{-}4)^0_0$
12. American	4	$(12, 5\text{-}8)^0_0$
13. Soviet	5	$(13, 10\text{-}11\text{-}4)^0_0$
14. Soviet	6	$(14, 4\text{-}10\text{-}11\text{-}13)^0_0$
15. Soviet	7	$(15, 4\text{-}10\text{-}11\text{-}13\text{-}14)^0_0$
16. Western	7	
17. Gorbachev	8	$(17, 6)^0_0$
18. Soviet	8	$(18, 4\text{-}10\text{-}11\text{-}13\text{-}14\text{-}15)^0_0$
19. U.S.S.R.	8	
20. Soviet	10	$(20, 4\text{-}10\text{-}11\text{-}13\text{-}14\text{-}15\text{-}18)^0_0$
21. Soviet	10	$(21, 4\text{-}10\text{-}11\text{-}13\text{-}14\text{-}15\text{-}18\text{-}20)^0_0$
22. U.S.S.R	11	$(22, 19)^0_0$
23. Communist Party	11	
24. Kommunist	11	$(24, 23)^0_0$
25. Soviet	13	$(25, 4\text{-}10\text{-}11\text{-}13\text{-}14\text{-}15\text{-}18\text{-}20\text{-}21)^0_0$
26. Soviet	15	$(26, 4\text{-}10\text{-}11\text{-}13\text{-}14\text{-}15\text{-}18\text{-}20\text{-}21\text{-}25)^0_0$
27. Politburo	15	
28. Krushchev	16	
29. Soviet	17	$(29, 4\text{-}10\text{-}11\text{-}13\text{-}14\text{-}15\text{-}18\text{-}20\text{-}21\text{-}25\text{-}26)^0_0$
30. Soviet	17	$(30, 4\text{-}10\text{-}11\text{-}13\text{-}14\text{-}15\text{-}18\text{-}20\text{-}21\text{-}25\text{-}26\text{-}29)^0_0$
31. Soviet	19	$(31, 4\text{-}10\text{-}11\text{-}13\text{-}14\text{-}15\text{-}18\text{-}20\text{-}21\text{-}25\text{-}26\text{-}29\text{-}30)^0_0$
32. United States	21	
33. Soviet Union	24	
34. Gorbachev	25	$(34, 6\text{-}17)^0_0$

Chain 1, Segment 2		
Word	Sentence	Lexical Chain
35. elite	31	

Chain 2 Segment 1		
Word	Sentence	Lexical Chain
1. woman	1	
2. woman	4	$(2, 1)_0^0$
3. husband	4	$(3, 2)_0^1$ $(3, 1)_0^1$
4. woman	4	$(4, 3)_0^1$ $(4, 2)_0^0$ $(4, 1)_0^0$
5. women	5	$(5, 4)_0^0$ $(5, 3)_0^1$ $(5, 2)_0^0$ $(5, 1)_0^0$
6. woman	6	$(6, 5)_0^0$ $(6, 4)_0^0$ $(6, 3)_0^1$ $(6, 2)_0^0$ $(6, 1)_0^0$
7. woman	7	$(7, 1\text{-}2\text{-}4\text{-}5\text{-}6)_0^0$ $(7, 3)_0^1$
8. women	8	$(8, 1\text{-}2\text{-}4\text{-}5\text{-}6\text{-}7)_0^0$ $(8, 3)_0^1$
9. women	8	$(9, 1\text{-}2\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8)_0^0$ $(9, 3)_0^1$
10. male	8	$(10, 1\text{-}2\text{-}3\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9)_0^2$
11. women	10	$(11, 1\text{-}2\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}10)_0^0$ $(11, 3\text{-}10)_0^2$
12. male	12	$(12, 10)_0^0$ $(12, 1\text{-}2\text{-}3\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}10\text{-}11)_0^2$
13. women	13	$(13, 3\text{-}10\text{-}12)_0^2$ $(13, 1\text{-}2\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}11)_0^0$
14. women	14	$(14, 3\text{-}10\text{-}12)_0^2$ $(14, 1\text{-}2\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}11\text{-}13)_0^0$
15. woman	15	$(15, 3\text{-}10\text{-}12)_0^2$ $(15, 1\text{-}2\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}11\text{-}13\text{-}14)_0^0$
16. women	17	$(16, 3\text{-}10\text{-}12)_0^2$ $(16, 1\text{-}2\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}11\text{-}13\text{-}14\text{-}15)_0^0$
17. woman	18	$(17, 3\text{-}10\text{-}12)_0^2$ $(17, 1\text{-}2\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}11\text{-}13\text{-}14\text{-}15\text{-}16)_0^0$
18. women	19	$(18, 3\text{-}10\text{-}12)_0^2$ $(18, 1\text{-}2\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}11\text{-}13\text{-}14\text{-}15\text{-}16\text{-}17)_0^0$
19. husband	20	$(19, 3)_0^0$ $(19, 1\text{-}2\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}10\text{-}11\text{-}12\text{-}13\text{-}14\text{-}15\text{-}16\text{-}17\text{-}18)_0^2$
20. wife	21	$(20, 1\text{-}2\text{-}3\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}10\text{-}11\text{-}12\text{-}13\text{-}14\text{-}15\text{-}16\text{-}17\text{-}18\text{-}19)_0^1$ $(10, 12)_0^2$
21. women	22	$(21, 3\text{-}10\text{-}12)_0^2$ $(21, 1\text{-}2\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}11\text{-}13\text{-}14\text{-}15\text{-}16\text{-}17)_0^0$ $(21, 19\text{-}20)_0^1$
22. women	24	$(22, 3\text{-}10\text{-}12)_0^2$ $(22, 1\text{-}2\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}11\text{-}13\text{-}14\text{-}15\text{-}16\text{-}17\text{-}21)_0^0$ $(22, 19\text{-}20)_0^1$
23. husband	25	$(23, 3\text{-}19)_0^0$ $(23, 1\text{-}2\text{-}4\text{-}5\text{-}6\text{-}7\text{-}8\text{-}9\text{-}10\text{-}11\text{-}12\text{-}13\text{-}14\text{-}15\text{-}16\text{-}17\text{-}18\text{-}20\text{-}21\text{-}22)_0^2$

Chain 2, Segment 2		
Word	Sentence	Lexical Chain
24. wives	31	(24, 1-2-3-4-5-6-7-8-9-10-11-12-13-14-15-16-17-18-19-21-22-23)$^{1}_{0}$ (24, 12)$^{2}_{0}$ (24, 20)$^{0}_{0}$
25. husband	31	(25, 3-19-23)$^{0}_{0}$ (25, 1-2-4-5-6-7-8-9-10-11-12-13-14-15-16-17-18-20-21-22-24)$^{2}_{0}$
26. wives	31	(26, 1-2-3-4-5-6-7-8-9-10-11-12-13-14-15-16-17-18-19-21-22-23-25)$^{1}_{0}$ (26, 12)$^{2}_{0}$ (26, 20-24)$^{0}_{0}$

Chain 3		
Word	Sentence	Lexical Chain
1. mother	21	
2. family	22	(2, 1)$^{1}_{0}$
3. home	24	(3, 1)$^{2}_{0}$ (3, 2)$^{1}_{0}$
4. family	24	(4, 3)$^{2}_{0}$ (4, 2)$^{0}_{0}$ (4, 1)$^{1}_{0}$
5. home	26	(5, 3)$^{0}_{0}$ (5, 4)$^{1}_{0}$ (5, 2)$^{1}_{0}$ (5, 1)$^{2}_{0}$

Chain 4, Segment 1		
Word	Sentence	Lexical Chain
1. image	7	
2. reality	8	(2, 1)$^{2}_{0}$
3. reality	10	(3, 2)$^{1}_{0}$ (3, 1)$^{2}_{0}$
4. real	12	(4, 2-3)$^{0}_{0}$ (4, 1)$^{2}_{0}$
5. real	14	(5, 2-3-4)$^{0}_{0}$ (5, 1)$^{2}_{0}$
6. theoretically	17	(6, 2-3-4-5)$^{4}_{0}$ (6, 1)$^{T2}_{1}$

Chain 4, Segment 2		
Word	Sentence	Lexical Chain
7. reality	24	(7, 6)$^{4}_{0}$ (7, 2-3-4-5)$^{0}_{0}$ (7, 1)$^{2}_{0}$

Chain 5		
Word	Sentence	Lexical Chain
1. summit	2	
2. summit	3	(2, 1)$^{0}_{0}$

Chain 6, Segment 1		
Word	Sentence	Lexical Chain
1. equal rights	9	
2. law	9	
3. propaganda	9	
4. law	10	$(4, 2)^{0}_{0}$
5. official	11	$(5, 1)^{1}_{0}$ $(5, 2-4)^{1}_{0}$
6. power	12	$(6, 5)^{1}_{0}$ $(6, 1-2-4)^{2}_{0}$
7. power	12	$(7, 6)^{0}_{0}$ $(7, 5)^{1}_{0}$ $(7, 1-2-4)^{2}_{0}$
8. policy	12	$(8, 6-7)^{5}_{0}$ $(8, 1-2-4-5)^{T7}_{1}$
9. power	13	$(9, 8)^{5}_{0}$ $(9, 6-7)^{0}_{0}$ $(9, 5)^{1}_{0}$ $(9, 1-2-4)^{2}_{0}$
10. bureaucracy	13	$(10, 2-4-5-6-7-9)^{1}_{0}$ $(10, 1-8)^{2}_{0}$
11. decision-making	14	
12. power	14	$(12, 8)^{5}_{0}$ $(12, 6-7)^{0}_{0}$ $(12, 5-10)^{1}_{0}$ $(12, 1-2-4)^{2}_{0}$
13. power	16	$(13, 8)^{5}_{0}$ $(13, 6-7-12)^{0}_{0}$ $(13, 5-10)^{1}_{0}$ $(13, 1-2-4)^{2}_{0}$
14. political	17	$(14, 2-4-5-6-7-9-10-12-13)^{1}_{0}$ $(14, 8)^{2}_{0}$ $(14, 1)^{T3}_{1}$
15. rights	17	$(15, 1)^{0}_{0}$
16. constitution	17	$(16, 1-15)^{1}_{0}$ $(16, 2-3-4-5-6-7-8-9-10-11-12-13-14)^{T1}_{1}$
17. legal	18	$(17, 1--15-16)^{1}_{0}$ $(17, 2-4-6-7-8-9-10-12-13-14)^{2}_{0}$

Chain 6, Segment 2		
Word	Sentence	Lexical Chain
18. power	25	$(18, 8)^{5}_{0}$ $(18, 6-7-12-13)^{0}_{0}$ $(18, 5-10)^{1}_{0}$ $(18, 1-2-4-14-15-17)^{2}_{0}$ $(18, 17)^{T1}_{1}$

Chain 6, Segment 3		
Word	Sentence	Lexical Chain
19. power	31	$(19, 8)^{5}_{0}$ $(19, 6-7-12-13-18)^{0}_{0}$ $(19, 5-10)^{1}_{0}$ $(19, 1-2-4-14-15-17)^{2}_{0}$ $(19, 17)^{T1}_{1}$

Chain 7		
Word	Sentence	Lexical Chain
1. work	19	
2. full-time	19	
3. housework	20	$(3, 1)_0^0$
4. full-time	20	$(4, 2)_0^0$
5. job	20	$(5, 1\text{-}3)_0^1$
6. salary	21	
7. working	21	$(7, 5)_0^1$ $(7, 1\text{-}3)_0^0$
8. work	22	$(8, 1\text{-}3\text{-}7)_0^0$ $(8, 5)_0^1$
9. money	22	$(9, 6)_0^2$
10. support	22	$(10, 1\text{-}3\text{-}7\text{-}8)_0^2$ $(10, 5)_1^{T8}$
11. salaries	23	$(11, 9)_0^2$ $(11, 6)_0^0$
12. job	25	$(12, 5)_0^0$ $(12, 1\text{-}3\text{-}7\text{-}8)_0^1$ $(12, 10)_1^{T8}$
13. full-time	25	$(13, 2\text{-}4)_0^0$
14. work	26	$(14, 1\text{-}3\text{-}7\text{-}8)_0^0$ $(14, 5\text{-}12)_0^1$ $(14, 10)_0^2$
15. worker	28	$(15, 1\text{-}3\text{-}7\text{-}8\text{-}14)_0^0$ $(15, 5\text{-}12)_0^1$ $(15, 10)_0^2$
16. wage	28	$(16, 6\text{-}11)_0^0$ $(16, 9)_0^2$
17. job	31	$(17, 5\text{-}12)_0^0$ $(17, 1\text{-}3\text{-}7\text{-}8\text{-}15)_0^1$ $(17, 10)_1^{T8}$

Chain 8, Segment 1		
Word	Sentence	Lexical Chain
1. academic	13	
2. academic	13	$(2, 1)_0^0$

Chain 8, Segment 2		
Word	Sentence	Lexical Chain
3. graduate study	27	$(3, 1\text{-}2)_0^2$
4. courses	29	$(4, 3)_0^2$ $(4, 1\text{-}2)_1^{T3}$
5. exams	29	$(5, 3)_0^1$ $(5, 4)_0^2$ $(5, 1\text{-}2)_1^{T3}$
6. exams	29	$(6, 5)_0^0$ $(6, 3)_0^1$ $(6, 4)_0^2$ $(6, 1\text{-}2)_1^{T3}$
7. dissertation	29	
8. papers	29	
9. study	30	$(9, 3)_0^0$ $(9, 4)_0^2$ $(9, 5\text{-}6)_0^1$ $(9, 1\text{-}2)_1^{T3}$
10. college	31	$(10, 1\text{-}2)_0^1$ $(10, 3\text{-}9)_0^2$ $(10, 5\text{-}6)_1^{T9}$ $(11, 4)_1^{T6}$

Chain 9		
Word	Sentence	Lexical Chain
1. role	3	
2. role	5	$(2, 1)_0^0$

4.5.3 The Intentional Structure

Figure 4.4 gives the intentional structure for this example.

Figure 4.3: The Intentional Structure of Example (4-4)

1 (1–31)
Describe why Raisa Gorvachev is not a realistic model of a typical Soviet woman's life.

- 1.1 (1–3)
  Describe the controversy and publicity that Mrs.‘Gorbachev stirred up during the summit meeting.
- 1.2 (4–17)
  Describe how her image is that of a role model for Soviet women, but in reality, she is not, and is not likely to be in the near future.
  - 1.2.1 (4–7)
    Describe Mrs.‘Gorvachev's image.
  - 1.2.2 (8–17)
    Describe the reality of life for women in the Soviet Union.
    - 1.2.2.1 (8–10)
      Describe why Mrs.‘Gorbachev's image is not reality.
    - 1.2.2.2 (11–17)
      Describe how life for women will not change in the near future, since power is a male preserve.
- 1.3 (18–31)
  Describe how and why women work in the Soviet Union, comparing this to what would be desirable.
  - 1.3.1 (18–26)
    Compare the average working woman's work situation with Mrs.‘Gorbachev's.
    - 1.3.1.1 (18–24)
      Describe that as is the case in the United States, it is hard to be both a working woman and a mother, but that the money is sometimes essential.
    - 1.3.1.2 (25–26)
      Compare the average situation to that of Mrs.‘Gorbachev, who got her part-time job due to the power status of her husband.
  - 1.3.2 (27–30)
    Describe how graduate study is the best life, since academic positions allow for one to two days a week of "sabbatical" time, which means "staying home."
  - 1.3.3 (31) Describe how powerful husbands with clout can get their wives academic positions.

4.5.4 The Correspondences between Lexical and Intentional Structures

The following table gives the correspondences between the lexical chains and intentions of example (4-4):

Chain	Chain Range	Intention	Intention Range
1.1	1–25	1	1–31
1.2	31	end of 1	31
2.1	1–25	1	1–31
2.2	31	end of 1	31
3	21–26	1.3.1	18–26
4.1	7–17	1.2.2	8–17
4.2	24	end of 1.3.1.1	24
5	2–3	1.1	1–3
6.1	9–18	1.2.2	8–17
6.2	25	start of 1.3.1.2	25
6.3	31	1.3.3	31
7	19–31	1.3	18–31
8	27–31	1.3.2	27–30
9	3–5	1.2.1	4–7

Chains 1 and 2 both span sentences 1 to 25, and then have a return in sentence 31, which is the last sentence of the example. The return corresponds to an intentional boundary, and the whole chain unifies the example. However, there is no intention that runs from sentences 1 to 25, and other textual information would be required to determine this.

Chain 3 corresponds to intention 1.3.1. Chain 4.1 corresponds to intention 1.2.2. Chain 4.2, which is a single word return to chain 4, corresponds to an intentional boundary—the end of intention 1.3.1.1. A chain return that corresponds to an intentional boundary is a regular occurence. However, unlike most returns, this one does not link lower-level intentions together into a higher-level one. Rather, it indicates semantic connections in the text that are not linked together by the intentional structure. The specific semantic connection is that of the reality of a woman's life and role in the Soviet Union. This concept runs through the article, but does not fit into the hierarchical intentional structure.

Chain 5 corresponds to intention 1.1. Chain 6.1 corresponds to intention 1.2.2. Chain 6.2 corresponds to an intentional boundary—the start of intention 1.3.1.2. Chain 6.3 corresponds to intention 1.3.3. However, as in the analysis of the returns in chain 4, these two returns do not indicate that there is one high-level chain running the entire length of the chain, from sentences 9 to 31. The returns are indicative of semantic connections in the text that are not indicated by the intentional structure. The semantic connection in this case is the idea that power can help your life. This idea runs throughout the entire text but is otherwise irrespective of the hierarchical structure.

Chain 7 corresponds to intention 1.3. Chain 8 corresponds to intention 1.3.2. Note that chain 8, segment 1, is not considered a part of chain 8 in this analysis. The algorithm (section 3.2.5) would have computed chain 8.2 as a return to chain 8.1, but this is not intuitive. Chain 9 corresponds to intention 1.2.1.

This example illustrates that there is a correspondence between lexical chains and intentions. The main difference in this example is that chain returns do not indicate a grouping of lower-level intentions into one higher-level intention spanning the entire chain. Rather, they indicate semantic connections in the text that are not indicated by the hierarchical intentional structure. The returns do, however, coincide with intentional boundaries, as happened in the other examples.

4.6 Example (4-5)

4.6.1 The Text

Here is the text of example (4-5), the first part of an article in *Equinox* magazine, September, 1987, by Adrian Forsyth, entitled "The Plague Within":[6]

(4-5)

1. ¶On the morning of June 21, 1630, Catarina Rosa stood terrified on her balcony as she watched a man smear a dark, gummy liquid on the walls of nearby buildings.
2. When she had recovered from her fear, Rosa gathered her neighbours, and they informed the Milan senate that an Annointer had been seen.
3. The search began for the spreader of the plague.
4. ¶After a series of false arrests and brutal beatings of innocent people, a minor health official, Guglielmo Piazza, was accused of annointing walls with a plague-causing ointment.
5. Piazza denied it.
6. His job was to record cases of plague, and he carried an ink horn on his belt.
7. He was just wiping ink from his fingers.
8. Torture changed his mind, causing him to implicate a barber, one Giangiacomo Mora, as the supplier of the ointment.
9. Mora also confessed after similar interrogation.
10. The court sentenced the men to death by extended torture.
11. ¶An ox-drawn wagon complete with a crew of priests and functionaries carried Piazza and Mora around the plaza before crowds of onlookers.
12. Officials heated large pincers until they were red-hot and repeatedly tore the flesh of the men.
13. With mallets and cleavers, they severed the right hand of each prisoner.
14. Next, the executioners stretched Piazza and Mora out on a platform in the plaza and broke their limbs with iron bars.

[6] ©1987 Adrian Forsyth. Reprinted by kind permission of the author.

15. Finally, they tied the two men to wagon wheels, hoisted them onto poles for six hours and then burned them to death.

16. ¶Piazza and Mora were, of course, innocent.

17. The plague, which ultimately killed 150,000 Milanese citizens, was not the simple work of Annointers but was the product of a complicated ecological interaction between rats, fleas, the bacterium *Yersinia pestis* and humans living in squalid conditions.

18. ¶Ironically, the importance of human poverty in the *ménage à quatre* was already known when the unfortunate Piazza and Mora were being carted around the plaza.

19. In nearby Florence, the health magistry had urged volunteers to improve the living quarters of the poor because "filth is the mother of the corruption of the air, and the latter is the mother of the plague."

20. Even the connection between rats and the plague was suspected by officials in European cities such as Frankfurt, and they attempted rat control by requiring Jews to pay a tax of thousands of rats' tails.

21. Improving urban ecology was obviously a known preventative of bubonic plague.

22. However, blaming Annointers was an easier and simpler reaction.

23. ¶It is tempting to believe that the attitudes which led to the persecution of Annointers are as far behind us as the epidemics of Black Death that ravaged Europe during the Middle Ages.

24. However, current attitudes toward the modern epidemic of AIDS (Aquired Immune Deficiency Syndrome) and the human immuno-deficiency virus (HIV) responsible for the condition show some unpleasant similarities to those during the bubonic plague.

25. Whether the epidemic is mediaeval bubonic plague or present-day AIDS, people often latch on to a simplistic explanation instead of searching for the ecological and medical origins of the epidemic.

26. In the past few years, influential Christian evangelists have broadcast the opinion that AIDS is divine retribution for homosexuality.

27. Even a medical journal asked editorially: "Might we be witnessing, in fact, in the form of a modern communicable disorder, a fulfillment of Saint Paul's pronouncement, 'The due penalty of their error'?"

28. ¶Such attitudes reflect a classic confusion of cause and effect that ignores the biology of infective diseases.

29. The HIV epidemic is not caused by homosexuality.

30. At a biological level, the association of AIDS with homosexual males, prostitutes, plasma recipients and intravenous drug users is a result not of moral weakness but simple of the high number of exposures to the bodily fluids of HIV carriers.

31. If HIV was able to spread through inhalation and exhalation, like a flu virus, then AIDS would be proliferating among groups such as symphony musicians, office workers, students and others who live and work in high densities.

4.6.2 The Lexical Structure

The following tables show the lexical chains found in example (4-5):

Chain 1		
Word	Sentence	Lexical Chain
1. terrified	1	
2. fear	2	(2, 1)$_0^1$
3. plague	3	(3, 1)$_0^1$ (3, 2)$_0^2$
4. brutal	4	(4, 2-3)$_0^2$ (4, 1)$_0^5$
5. plague	4	(5, 3)$_0^0$ (5, 1)$_0^1$ (5, 2-4)$_0^2$
6. plague	6	(6, 3-5)$_0^0$ (6, 1)$_0^1$ (6, 2-4)$_0^2$
7. torture	8	(7, 1-3-4-5-6)$_0^1$ (7, 2)$_0^2$
8. death	10	(8, 3-5-6-7)$_0^2$ (8, 1-2-4)$_1^{T6}$
9. torture	10	(9, 7)$_0^0$ (9, 1-3-4-5-6-8)$_0^1$ (9, 2)$_0^2$
10. death	15	(10, 8)$_0^0$ (10, 9)$_0^1$ (10, 3-5-6-7)$_0^2$ (10, 1-2-4)$_1^{T6}$
11. plague	17	(11, 3-5-6)$_0^0$ (11, 1-7-9)$_0^1$ (11, 4-8-9)$_0^2$
12. plague	19	(12, 3-5-6-11)$_0^0$ (12, 1-7-9)$_0^1$ (12, 4-8-9)$_0^2$
13. plague	20	(13, 3-5-6-11-12)$_0^0$ (13, 1-7-9)$_0^1$ (13, 4-8-9)$_0^2$
14. plague	21	(14, 3-5-6-11-12-13)$_0^0$ (14, 1-7-9)$_0^1$ (14, 4-8-9)$_0^2$
15. black death	23	
16. epidemic	24	(16, 3-5-6-11-12-13-14)$_0^1$ (16, 1-2-4-7-8-9-10)$_1^{T10}$
17. plague	24	(17, 3-5-6-11-12-13-14)$_0^0$ (17, 1-7-9-16)$_0^1$ (17, 4-8-9)$_0^2$
18. plague	25	(18, 3-5-6-11-12-13-14-17)$_0^0$ (18, 1-7-9-16)$_0^1$ (18, 4-8-9)$_0^2$
19. epidemic	25	(19, 3-5-6-11-12-13-14-17-18)$_0^1$ (19, 16)$_0^0$ (19, 1-2-4-7-8-9-10)$_1^{T10}$
20. diseases	28	(20, 3-5-6-11-12-13-14-17-18)$_0^1$ (20, 16-19)$_0^0$ (20, 1-2-4-7-8-9-10)$_1^{T10}$
21. epidemic	29	(21, 3-5-6-11-12-13-14-17-18-20)$_0^1$ (21, 16-19)$_0^0$ (21, 1-2-4-7-8-9-10)$_1^{T10}$
22. virus	31	(22, 3-5-6-7-9-11-12-13-14-16-17-18-19-20)$_0^2$ (22, 1-2-4-8-10)$_1^{T10}$

Chain 2		
Word	Sentence	Lexical Chain
1. search	3	
2. arrests	4	
3. innocent	4	
4. accused	4	$(4, 2)^{1}_{0}$
5. denied	5	
6. implicate	8	$(6, 2\text{-}4)^{2}_{0}$
7. confessed	9	
8. interrogation	9	$(8, 1)^{1}_{0}$
9. court	10	$(9, 1\text{-}5\text{-}8)^{2}_{0}$
10. sentenced	10	$(10, 2\text{-}4\text{-}6\text{-}7)^{2}_{0}$
11. executioner	14	
12. innocent	16	$(12, 3)^{0}_{0}$

Chain 3		
Word	Sentence	Lexical Chain
1. wagon	11	
2. wagon	15	$(2, 1)^{0}_{0}$
3. carted	18	$(3, 1\text{-}2)^{1}_{0}$

Chain 4		
Word	Sentence	Lexical Chain
1. flesh	12	
2. hand	13	$(2, 1)^{1}_{0}$
3. limbs	14	$(3, 2)^{2}_{0}$ $(3, 1)^{T2}_{1}$

Chain 5		
Word	Sentence	Lexical Chain
1. rats	17	
2. fleas	17	
3. pest	17	
4. squalid	17	$(4, 2)^{1}_{0}$
5. poverty	18	
6. poor	19	$(6, 5)^{1}_{0}$ $(6, 3)^{2}_{0}$
7. filth	19	$(7, 4)^{1}_{0}$ $(7, 2)^{1}_{0}$
8. corruption	19	$(8, 2\text{-}4\text{-}7)^{1}_{0}$
9. rats	20	$(9, 1)^{0}_{0}$
10. rats	20	$(10, 1\text{-}9)^{0}_{0}$

Chain 6		
Word	Sentence	Lexical Chain
1. Florence	19	
2. European	20	
3. cities	20	
4. Frankfurt	20	
5. Europe	23	$(5, 2)^{0}_{0}$

Chain 7		
Word	Sentence	Lexical Chain
1. AIDS	24	
2. HIV	24	
3. AIDS	25	$(3, 1)_0^0$
4. AIDS	26	$(4, 1\text{-}3)_0^0$
5. HIV	29	$(5, 2)_0^0$
6. homosexuality	29	
7. HIV	30	$(7, 2\text{-}5)_0^0$
8. AIDS	30	$(8, 1\text{-}3\text{-}4)_0^0$
9. homosexual	30	$(9, 6)_0^0$
10. prostitutes	30	
11. intravenous	30	
12. drugs	30	
13. HIV	31	$(13, 2\text{-}5\text{-}7)_0^0$
14. AIDS	31	$(14, 1\text{-}3\text{-}4\text{-}8)_0^0$

Chain 8		
Word	Sentence	Lexical Chain
1. Christion	26	
2. Evangelists	26	$(2, 1)_0^1$
3. divine	26	$(3, 1)_0^2$ $(3, 2)_1^{T2}$
4. retribution	26	
5. St. Paul's	27	
6. moral	30	$(6, 3)_0^2$ $(6, 1\text{-}2)_1^{T3}$

Chain 9		
Word	Sentence	Lexical Chain
1. tore	12	
2. severed	13	$(2, 1)_0^1$
3. broke	14	$(3, 1\text{-}2)_0^1$

Chain 10		
Word	Sentence	Lexical Chain
1. Rosa	1	
2. Rosa	2	$(2, 1)_0^0$

Chain 11		
Word	Sentence	Lexical Chain
1. Piazza	4	
2. Mora	8	
3. Piazza	10	$(3, 1)_0^0$
4. Mora	10	$(4, 2)_0^0$
5. Piazza	14	$(5, 1\text{-}3)_0^0$
6. Mora	14	$(6, 2\text{-}4)_0^0$
7. Piazza	16	$(7, 1\text{-}3\text{-}5)_0^0$
8. Mora	16	$(8, 2\text{-}4\text{-}6)_0^0$
9. Piazza	17	$(9, 1\text{-}3\text{-}5\text{-}7)_0^0$
10. Mora	17	$(10, 2\text{-}4\text{-}6\text{-}8)_0^0$
11. Piazza	18	$(11, 1\text{-}3\text{-}5\text{-}7\text{-}9)_0^0$
12. Mora	18	$(12, 2\text{-}4\text{-}6\text{-}8\text{-}10)_0^0$

Chain 12, Segment 1		
Word	Sentence	Lexical Chain
1. Annointer	2	
2. annointing	4	$(2, 1)_0^0$

Chain 12, Segment 2		
Word	Sentence	Lexical Chain
3. Annointer	17	$(3, 1\text{-}2)_0^0$

Chain 12, Segment 3		
Word	Sentence	Lexical Chain
4. Annointers	22	$(4, 1\text{-}2\text{-}3)_0^0$
5. Annointers	23	$(5, 1\text{-}2\text{-}3\text{-}4)_0^0$

Chain 13		
Word	Sentence	Lexical Chain
1. attitudes	23	
2. attitudes	24	$(2, 1)_0^0$
3. attitudes	28	$(3, 1\text{-}2)_0^0$

Chain 14		
Word	Sentence	Lexical Chain
1. mother	20	
2. mother	20	$(2, 1)_0^0$

4.6.3 The Intentional Structure

Figure 4.5 gives the intentional structure for this example.

Figure 4.4: The Intentional Structure of Example (4-5)

1 (1–31)
Describe the similarities between the false annointer explanation of the Black Death, and the false Christian moral punishment explanation explanation of AIDS.

- 1.1 (1–22)
  Describe the situation of the Annointers.
  - 1.1.1 (1–15)
    Describe the annointer explanation of Black Death.
    - 1.1.1.1 (1–10)
      Describe the false arrest of Piazza and Mora.
    - 1.1.1.2 (11–15)
      Describe the subsequent torture of Piazza and Mora.
  - 1.1.2 (16–22)
    Describe the real reason for Black Death which involved poverty, rats, and fleas.
- 1.2 (23–31)
  Compare the Annointers and Plague scenario with the divine retribution and AIDS scenario.
  - 1.2.1 (23–25)
    Describe how people use a simplistic explanation rather that a scientific one.
  - 1.2.2 (26–27)
    Describe the false Christian moral explanation for AIDS.
  - 1.2.3 (28–31)
    Describe the real medical explanation for AIDS.

4.6.4 The Correspondences between Lexical and Intentional Structures

The following table gives the correspondences between the lexical chains and intentions of example (4-5):

Chain	Chain Range	Intention	Intention Range
1	1–31	1	1–31
2	3–16	1.1.1	1–15
3	11–18	1.1.1.2	11–15
4	12–14	1.1.1.2	11–15
5	17–20	1.1.2	16–22
6	19–23	1.1.2	16–22
7	24–31	1.2	23–31
8	26–30	1.2.2 and 1.2.3	26–31
9	12–14	1.1.1.2	11–15
10	1–2		
11	4–18	1.1.1	1–15
12.1	2–4	part of 1.1	
12.2	17	start of 1.1.2	16
12.3	22–23	end of 1.1	22
13	23–28	1.2	23–31
14	20		

Chain 1 corresponds to intention 1 since they both span the entire example. Chain 2 corresponds to intention 1.1.1. Chain 3 corresponds to intention 1.1.1.2. In this case the chain runs three sentences longer than its corresponding intention, which is rare. In most cases, even though the chain does not match the intention exactly, the chain's boundaries are within the boundaries of the intention. Chain 4 also corresponds to intention 1.1.1.2. These two chains provide a good illustration of the fact that the chains do not, on their own, provide conclusive information about the intentional structure. Chain 4, for example, is a chain within the boundaries of chain 3, but does not correspond to a low-level intention inside intention 1.1.1.2.

Chains 5 and 6 both correspond to intention 1.1.2. In this case, there are two chains corresponding to one intention, and neither of them do so exactly. The chains suggest the presence of an intentional component in their vicinity.

Chain 7 corresponds to intention 1.2. Chain 8 spans the sentences of two intentions: 1.2.2, and 1.2.3. Other textual information such as semantics or pragmatics would have to be used to properly separate the two intentions.

Chain 9 corresponds to intention 1.1.1.2. Chain 10 (a two-word chain) does not have a corresponding intention in my analysis, however with a finer granularity of intentional structure analysis it might. Chain 11 corresponds to intention 1.1.1. This is another rare example (see chain 3) of a chain going past its corresponding intention.

Chain 12 consists of three short segments, which gives an indication that there is a structural unity in the sentences spanned by the entire chain. The three chain segments collectively span the sentences of intention 1.1. Each individual segment does not have a corresponding intention. Segment 12.1 has no corresponding intention, segment 12.2 occurs at the start of intention 1.1.1.2, and segment 12.3 occurs at the end of intention 1.1. The chain returns do, therefore, correspond to intentional boundaries.

Chain 13 corresponds to intention 1.2. Chain 14 does not correspond to an intention; however, it is simply a two-word reiteration spanning only one sentence.

Example (4-5) illustrates that the lexical chain information is useful as an indicator of structure, but that there is not a one-to-one mapping between chains and intentions. In two cases (chains 3 and 4, and chains 5 and 6) two overlapping chains corresponded to the same intention. In one case (chain 8) one chain corresponded to two intentions. There were two cases of chain boundaries exceeding the boundaries of their corresponding intentions. These cases show the necessity of integrating this tool with other sources of textual information. Chain returns were useful as intentional boundary indicators, and in unifying low-level intentions into a high-level intention.

4.7 Conclusions

The five examples given in this chapter show that the lexical chains computed by the algorithm in section 3.2.5 are useful as an indicator of the intentional structure proposed by Grosz and Sidner [4] (detailed in section 1.4.2). A major problem with using their theory was that there was no way of computing the intentions or linguistic segments.

Very few lexical chains (three) had no correspondence to an intention. All three of the chains that did not have a correspondence were two- or three-word chains that spanned only one or two sentences. In terms of chain strength, they would be considered weak, and therefore less likely to be related to text structure. Perhaps with a finer granularity of intentional structure analysis they would correspond to a very low-level intention.

Almost all of the lexical chains had a corresponding intention. However some of the correspondences are more exact (in terms of spanning the same sentences) than others as the following table shows:

Example	Total Number of Correspondences	Number of Exact Correspondences	Number of Correspondences Within Two Sentences
1	10	3	5
2	27	8	11
3	8	3	4
4	14	5	6
5	14	1	7

There are a significant number of close, and hence easily computable, correspondences between the lexical chains and intentions. For the rest of the cases, an integration with more textual information is required to determine or validate the correspondences that exist.

Chain returns were found to be indicative of intentional boundaries. They were also shown to be useful in tying together lower-level intentions into a higher-level intention. Two clear examples of this are example (4-1), chain 2, and example (4-3), chain 2. In example (4-4), the returns to chains 4 and 6 were unique in that they did not unify lower-level intentions into a higher-level intention. Rather, these returns indicated semantic connectivity between sentences in the text that did not fit into the hierarchical intentional structure.

Chain strength was used in a couple of cases as an indicator of text structure. In example (4-1), chain 1, *population* was repeated many times in a specific portion of the chain deemed strong because of this. This strong portion of the chain corresponded with an intention. Much more work needs to be done in the area of chain strength analysis.

Chapter 5

Conclusions

The motivation behind this work was that lexical cohesion in text should correspond in some way to the structure of the text. Knowledge of the structure of text is essential in determining its "deep" meaning. Since lexical cohesion is a result of a unit of text being, in some recognizable semantic way, about the same thing, and text structure analysis involves finding the units of text that are about the same thing, one should have something to say about the other. This was found to be true. As detailed in chapter 4, the lexical chains computed by the algorithm given in section 3.2.5 correspond closely to the intentional structure produced from the structural analysis method of Grosz and Sidner [4]. This is important, since Grosz and Sidner give no method for computing the intentions or linguistic segments that make up the structure that they propose.

Hence, the concept of lexical cohesion, defined originally by Halliday and Hasan [5] and expanded in this work, has a definite use in an automated text understanding system. The lexical chains are shown to be almost entirely computable with the aid of an on-line thesaurus containing the lookup methods outlined in section 3.2.2.1. The computer implementation of this type of thesaurus access would be a straightforward task involving traditional software engineering methodology and data bases. Writing the program to implement the algorithm given in section 3.2.5 would also be a straightforward task.

The examples used in this thesis are general-interest articles taken from five magazines. These examples were chosen specifically to illustrate that lexical cohesion, and hence this tool, are not domain-specific.

5.1 Further Research

It has already been mentioned that the concept of chain strength needs much further work. The intuition is that the stronger a chain, the more likely it is to have a corresponding structural component.

The question of how closely, if at all, lexical chains correspond to paragraph boundaries has not been studied here. The chains found in the examples given in chapter 4 are not contained within paragraphs, but there may, in fact, be some correspondence between boundaries of paragraphs and boundaries of lexical chains.

The integration of this tool with other text understanding tools is an area that will require a lot of work. Lexical chains do not always correspond exactly to intentional structure, and when they do not, other textual information is needed to obtain the correct correspondences. In the examples given, there were cases where a lexical chain did correspond to an intention, but the sentences spanned by the lexical chain and the intention differed by more than two. This could happen, for example, because the lexical chain started more than two sentences earlier than the intention, or if the lexical chain ended more than two sentences later than the intention. In these cases, verification of the possible correspondence must be accomplished through the use of other textual information such as semantics, pragmatics, or structure.

Another example of the need to integrate this tool with other sources of textual information comes from example (4-2), where three short overlapping chains, chains 7, 8, and 9 all correspond together to one intention spanning the sentences spanned collectively by all three chains. It turns out that chains 7, 8, and 9 are all related in the meaning context of the text, and this fact could be used to join the three chains together to correspond to one single intention (see section 4.3.4 for details).

It would be useful and straightforward to automate this tool and run a large corpus of text through it. I suspect that the chain-forming parameter settings (regarding transitivity and distances between words) will be shown to vary slightly according to author's style and the type of text. As it is impossible to do a complete and error-free lexical analysis of large text examples in a limited time-frame, automation is desirable.

A practical limitation of this work is that it depends on a thesaurus as its knowledge base. A thesaurus is as good as the work that went into creating it, and also depends on the perceptions, experience, and knowledge of its creators. Since language is not static, a thesaurus would have to be continually updated to remain current. Furthermore, no one thesaurus exists that meets all needs. *Roget's Thesaurus*, for example, is a general thesaurus that does not contain lexical relations specific to the geography of Africa or quantum mechanics. Therefore, further work needs to be done on identifying other sources of word knowledge such as domain specific thesauri, dictionaries, and statistical word usage information, that should be integrated with this work.

Qualification should be done of when word relationships cannot be determined from a knowledge source such as a thesaurus. An example of a problem area is metaphor. Consider this example [1]:

(5-1) 1. The moon is like a paper rose.

Moon and *rose* are lexically related only in the specific context of the metaphor.

Chapter 1 mentioned that lexical chains would be useful in providing a context for word sense disambiguation and in narrowing to specific word meanings. As an example of a chain providing useful information for word sense disambiguation, consider example (4-1), chain 2.1, words 1 to 15: {*afflicted, darkness, panicky, mournful, exciting, deadly, hating, aversion, cruel, relentless, weird, eerie, cold, barren, sterile* ...}. In the context of all of these words, it is clear that *barren* and *sterile* do not refer to the inability to reproduce, but to a *cruel cold*ness. As an example of a chain providing useful information for narrowing to specific word meanings, consider example (4-5), chain 5, words 1 to 7: {*rats, fleas, pest, squalid, poverty, poor, filth* ...}. By using the context provided by the

[1] Ian Lancashire, personal communication.

chain, it is clear that *filth* means more than just dirty. It means the *pest*-causing dirt that is a result of human *poverty.* This point requires further work as it has not been expanded on here.

Bibliography

[1] Bryan, Robert M. (1973), "Abstract Thesauri and Graph Theory Applications to Thesaurus Research," in Sally Yeates Sedelow, ed., *Automated Language Analysis.* 1973. University of Kansas, Departments of Computer Science and Linguistics.

[2] Carroll, Lewis (1960), *The Annotated Alice: Alice's Adventures in Wonderland and Through the Looking Glass.* New York: C.N. Potter.

[3] Danes, F. (1974) "Functional Sentence Perspective and the Organization of the Text." In: *Papers on Functional Sentence Perspective.* Prague: Academia, p. 106–128.

[4] Grosz, B., and Sidner, C. (1986). "Attention, Intentions and the Structure of Discourse". *Computational Linguistics* Volume 12, Number 3, p. 175–204.

[5] Halliday, Michael, and Hasan, Ruqaiya (1976). *Cohesion in English.* London: Longman Group.

[6] Hahn, Udo (1985). "On Lexically Distributed Text Parsing. A Computational Model for the Analysis of Textuality on the Level of Text Cohesion and Text Coherence." Universitat Konstanz. In: F. Kiefer (ed): *Linking in Text.*

[7] Hahn, Udo, and Reimer, Ulrich (1984). "Computing Text Consistency: An Algorithmic Approach to the Generation of Text Graphs". Universitat Konstanz. In: C. J. van Rijsbergen (ed): *Research and Development in Information Retrieval.* Proceedings of the 3rd Joint BCS and ACM Symposium. King's College, Cambridge, England, 2–6 July 1984. Cambridge: Cambridge U.P., 1984, p. 343–368.

[8] Hirst, G. (1981) *Anaphora in Natural Language Understanding: A Survey.* Lecture Notes in Computer Science. Berlin: Springer Verlag.

[9] Hirst, G. (1987). *Semantic interpretation and the resolution of ambiguity.* Studies in Natural Language Processing. Cambridge, England: Cambridge University Press.

[10] Hobbs, J. (1978). "Coherence and Coreference". Technical note 168, SRI International.

[11] Ide, N. (1987). "The Computational Approach to Meaning in Literary Texts". Department of Computer Science, Vassar College, Poughkeepsie N.Y.

[12] Lancashire, Ian. (1987). "Using a Textbase for English Language Research". In: *The Users of Large Text Databases.* Waterloo, Ontario: The University of Waterloo Centre for the New Oxford English Dictionary, p. 51–64.

[13] Lancashire, Ian (1987). Review of [15]. *Computational Linguistics*, Volume 13, Numbers 3–4, p. 347–350.

[14] McKeown, K. (1985). *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text.* Studies in Natural Language Processing. Cambridge England: Cambridge University Press.

[15] Phillips, M. (1984). *Aspects of Text Structure.* Linguistic series 52. Amsterdam: North-Holland.

[16] Postman, Leo, and Keppel, Geoffrey (editors) (1970). *Norms of Word Association.* New York: Academic Press.

[17] Reichman, R. (1985). *Getting Computers to Talk Like You and Me: Discourse Context, Focus, and Semantics (An ATN Model).* Cambridge, Massachusetts: The MIT Press.

[18] Roget, P. (1953). *Roget's Thesaurus, 3rd Edition.* Penguin Books Ltd.

[19] Roget, P. (1977). *Roget's International Thesaurus, Fourth Edition.* Harper and Row Publishers Inc.

[20] Sedelow, Sally, and Sedelow, Walter (1986). "Thesaural Knowledge Representation". Proceedings of the 2nd Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary: Advances in Lexicology. University of Waterloo.

[21] Sedelow, Sally, and Sedelow, Walter, (1987). "Semantic Space". *Computers and Translation* Volume 2, p. 235–245.

[22] Ventola, E. (1987). *The Structure of Social Interaction: A Systemic Approach to the Semiotics of Service Encounters.* Open Linguistic Series. London: Frances Pinter Publishers.