

# Experimental application of semantic clusters in WordNet

YunXiao Zhao

A Major Report  
in  
The Department  
of  
Computer Science

Presented in Partial Fulfillment of the Requirements  
for the Degree of Master of Computer Science at

Concordia University  
Montreal, Quebec, Canada

December 2003

@YunXiao Zhao, 2003



National Library  
of Canada

Bibliothèque nationale  
du Canada

Acquisitions and  
Bibliographic Services

Acquisitions et  
services bibliographiques

395 Wellington Street  
Ottawa ON K1A 0N4  
Canada

395, rue Wellington  
Ottawa ON K1A 0N4  
Canada

*Your file    Votre référence*

*ISBN: 0-612-91159-4*

*Our file    Notre référence*

*ISBN: 0-612-91159-4*

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

---

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this dissertation.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de ce manuscrit.

While these forms may be included in the document page count, their removal does not represent any loss of content from the dissertation.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

**Canada**

## **Acknowledgements**

I would like to take this opportunity to express my sincere thanks to my supervisor, Dr. Bergler, Sabine. Her conscientious supervision, her creative ideas, her constructive advice, and her erudite knowledge in Computer Science sustained me throughout the development of this major report. I will give my great thanks to the reader of my major report, Dr Kosseim. With her rigorous requests and advices, this major report is more precise.

I also wish to thank the CLaC colleagues Dr. Rene Witte, Frank Rudzicz, Michelle Khalife, and Zhuoyan Li, for their help especially with Scheme.

Furthermore, thanks to all the professors and staff in the department of computer science, especially Halina Monkiewicz, for their excellent support during my studies.

And also I have to thank my Chinese friend ShiQing, Zhao for his help in finishing my project documentation.

Finally, I would like to thank my wife, two sons and good friends for their love, encouragement, and support.

## Contents

<b>1. INTRODUCTION.....</b>	<b>3</b>
<b>2. LITERATURE OVERVIEW.....</b>	<b>4</b>
2.1. WHY WORDNET .....	6
2.2. CONCEPTS .....	9
2.2.1. <i>Similarity</i> .....	9
2.2.2. <i>Classificaton</i> .....	9
2.2.3. <i>Semantic Relatedness</i> .....	10
<b>3. METHODOLOGY.....</b>	<b>12</b>
3.1. DEFINITIONS .....	14
3.2. ALGORITHM .....	16
3.2.2. <i>Tables</i> .....	17
3.2.3. <i>Process stages</i> .....	22
3.3. SCHEME CODE.....	26
3.3.1. <i>Gambit</i> .....	26
3.3.2. <i>Scheme Code</i> .....	28
3.3.3. <i>Performance Interface</i> .....	29
<b>4. RESULTS AND EVALUATIONS.....</b>	<b>31</b>
4.1. RESULTS AND ANALYSIS OF SINGLE DOCUMENT CLUSTERING .....	32
4.1.1. <i>Sample-1 (Entertainment-AP880328-0206)</i> .....	33
4.1.2. <i>Sample-2 (Entertainment-AP891110-0043)</i> .....	36
4.1.3. <i>Sample-3 (Natural Disaster-AP890925-0054)</i> .....	37
4.1.4. <i>Sample-4 (Entertainment-AP900220-0065)</i> .....	37
4.1.5. <i>Sample-5 (Economy-FT923-5797.tagged)</i> .....	38
4.1.6. <i>Sample-6 (Economy-FT923-5835.tagged)</i> .....	39
4.1.7. <i>Combination Cluster</i> .....	40
4.2. DOCUMENTS CLOSENESS ANALYSIS .....	42
<b>5. CONCLUSION AND FUTURE WORK .....</b>	<b>49</b>
<b>REFERENCES .....</b>	<b>50</b>
<b>APPENDIX A. SCHEME CODE SAMPLES .....</b>	<b>52</b>
<b>APPENDIX B. CONTENTS OF TAGGED FILES .....</b>	<b>56</b>

# **Experimental application of semantic clusters in WordNet**

## **Abstract**

This paper reports on experiments into the semantics of nouns in documents through WordNet. Because WordNet is founded on semantic relations, we can compare every two nouns of a document with their common categories to discover whether they have semantic relations and then group them into a cluster if they do. In this project, generally we just verify three semantic relations of WordNet, namely hypernym, meronym and synonym.

Assume once we have the clusters, we can tell the basic meanings of a document. Based on this assumption, we can compare those documents by comparing their clusters to see their closure.

This project is designed on the above purpose and implemented under Unix system using the Scheme language.

# 1. Introduction

It is well known that the easiest measure to compare two words is simply to check their similarity. But for their internal semantic relations between words, we have to compare their semantic relatedness, extending to the comparison of documents.

Considering multiple senses for a word, in shallow processing applications like semantic pre-processing for document categorization it will be sufficient to use an underspecified sense instead of needless disambiguation between senses that are roughly equal in their relevance to a certain document category. Similarly, in shallow syntactic processing tasks, like statistical disambiguation of PP-attachment, the use of underspecified senses may be preferable as shown in experiments by {Krymolowski and Roth 1998}.

This project is first to cluster all the related nouns of a document by their semantic relatedness in WordNet, of a computational lexicon. Three measures (hypernym, synonym and meronym) of semantic relatedness are used, and the clusters of the three measures will be overlapped. With this step's result, we will verify their links in a cluster and tell: Do they really have relationship? If yes, which? If no, why?

Based on the clusters generated for each single document, we can treat them as approximative meanings for each document and compare the documents using the same measure as the first step. Once we get the clusters on this step, we will compare each single document's clusters with the combined final clusters, and we can get two values of their comparison.

In the final step, we will compare those two variations between two documents and get the optimized value to express their closeness of each two documents.

And this project will accept multiple documents to compare their closeness between each two documents in a final report.

## 2. Literature Overview

### 2.1. Word clustering

All types of word clustering reviewed have been widely studied from the theoretical point of view in the field of lexical semantics where it is commonly assumed that the semantic properties of a lexical item are fully reflected in the relations it keeps actual and potential linguistic contexts, namely on the syntagmatic and paradigmatic axes.

According to the EAGLES preliminary report [The EAGLES Lexicon Interest Group. 1999], *“Word clustering is a technique for partitioning sets of words into subsets of semantically similar words and is increasingly becoming a major technique used in a number of NLP tasks ranging from word sense or structural disambiguation to information retrieval and filtering.”*

In the literature, the EAGLES research group declared that two main different types of similarity have been used which can be characterised as follows:

1. **Paradigmatic**, or substitutional, similarity: two words that are paradigmatically similar may be substituted for one another in a particular context. For example, in the context *I read the book*, the word *book* can be replaced by *magazine* with no violation of the semantic well-

formedness of the sentence, and therefore the two words can be said to be paradigmatically similar;

2. **Syntagmatic** similarity: two words that are syntagmatically similar significantly occur together in 2 texts. For instance, *cut* and *knife* are syntagmatically similar since they typically co-occur within the same context.

Both types of similarity, computed through different methods, are used in the framework of a wide range of NLP applications.

Paradigmatic relations such as synonymy, hyperonymy/hyponymy, antonymy or meronymy occupy focal positions in discussions of lexical semantics. They reflect the way reality is categorised, subcategorised and graded along specific dimensions of lexical variation. By contrast, syntagmatic aspects of lexical meanings form a less prominent topic of lexical semantics which in the literature is generally referred to as co-occurrence restrictions.

## **2.2. NLP applications using word clustering techniques**

The various word clustering techniques have a large number of potentially important applications, as the EAGLES group described [The EAGLES Lexicon Interest Group. 1999], including:

- helping lexicographers in identifying normal and conventional usage;
- helping computational linguists in compiling lexicons with lexico-semantic knowledge;
- providing disambiguation cues for:



1. parsing highly ambiguous syntactic structures (such as noun compounds, complex coordinated structures, complement attachment, subject/object assignment for languages like Italian);
2. sense identification;
  - retrieving texts and/or information from large databases;
  - constraining the language model for speech recognition and optical character recognition (to help disambiguating among phonetically or optically confusable words).

Stand-alone applications require lexical semantic information. In shallow processing applications like semantic pre-processing for document categorization, it will be sufficient to use an underspecified sense instead of needless disambiguation between senses that are roughly equal in their relevance to a certain document category.

In this project we treat semantic similarity and historical straightforward similarity as one as relatedness, we just use shallow semantics to cluster all kinds of related noun words.

## **2.1. Why WordNet**

As Dr. Bergler (CS Dept. Concordia University) mentioned in her paper [Sabine Bergler 1993], *“a perceived need for computational lexica, or lexical databases that consistent, structured, striving for comprehensiveness but admitting that completeness is not achievable. In order to build these lexica, conventional dictionaries will have to be mined for their information.”*

WordNet, a lexica database built at Princeton University can be described as: “*WordNet is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept.*” [Miller, G. 1990]

In WordNet, different relations link synonym sets. For example, two synonym sets can be connected by a Hypernym link, which indicates that the words in the source synonym set are instances of the words in the target synonym set.

WordNet distinguishes between semantic and lexical relations, where the first ones are relations between meanings such as hypernym/hyponym, meronym, while the second ones are relations between words such as synonym.

Hyponymy/hypernymy (also indicated with subordination/superordination, subset/superset, or the ISA relation) is a semantic relation between word meanings. Hypernym is the generic term used to designate a whole class of instances. Y is a hypernym of X if X is a (kind of) Y.

For example, an apple is a kind of edible fruit, so edible fruit is a hypernym of apple and apple is a hyponym of edible fruit. Only nouns and verbs in WordNet participate in the hypernym relation.

Synonymy is, of course, a lexical relation between word forms. On the other hand, it should be considered the most important relation among existing ones. Actually since the ability to judge

that relation between word forms is a prerequisite for the representation of meanings in a lexical matrix, it is clear that the most important relation for WordNet is similarity of meaning. This relation is assumed to be a symmetric one, i.e. if x is similar to y then y is equally similar to x. For example, 'car' and 'wheeled vehicle', they are similar and they are equal to the same meaning. So we say these two words have a synonym relation.

Meronymy is the part/whole or HASA relation, known as meronym/holonym. The name of a constituent part of, the substance of, member of something. X is a meronym of Y if X is a part of Y. For example, the noun 'car' has part 'bumper', so we say 'bumper' is meronym of 'car'.

By means of these relations, all meanings can be interconnected, constituting a huge network or wordnet. Such a wordnet can be used for making various semantic inferences about the meanings of words (e.g. what words can name diseases), for finding alternative expressions or wordings, or for simply expanding words to sets of semantically related or close words in information retrieval. This approach has been developed in Princeton by G.A. Miller and his colleagues (Miller et al. 1991), and its latest version is known as WordNet 1.7.1.

We use the features of semantic relations in WordNet to group the words with semantic relations of documents into corresponding clusters.

## 2.2 Concepts

In this section, we need to briefly describe several very important concepts and relationships between them, especially about semantic relatedness used in this project.

### 2.2.1. Similarity

Similarity is an important and widely used concept. Many similarity measures have been proposed. Dekang Lin (CS Dept. University Manitoba) summarized and presented an information theoretic definition of similarity [D.Lin, 1998]. Intuition 1: The similarity between A and B is related to their commonality. The more commonality they share, the more similar they are. Intuition 2: The similarity between A and B is related to the differences between them. The more differences they have, the less similar they are. Intuition 3: The maximum similarity between A and B is reached when A and B are identical, no matter how much commonality they share. The similarity measure is not defined directly by a formula, rather, it is derived from a set of assumptions about similarity. In other words, if the assumptions are deemed reasonable, the similarity measure necessarily follows.

This definition of similarity is satisfied with our project's fundamental theory. We connect similarity to semantic relatedness to make our project performance theoretically reasonable.

### 2.2.2. Classification

In general, *“a classification is a method for organizing information. Human beings classify things spontaneously, a classification groups similar things together”* [Gary Olsen, 2003].

This definition is necessarily vague as he mentioned, there are many reasonable ways of defining similarity, and hence many alternative classifications for the same things. It looks like there is no "right way" to classify things. For example, words can be classified as: by their spelling (as in a dictionary); by the spelling of a foreign language synonym (as in a bilingual dictionary); by their meanings (as in a thesaurus). But the intended use of a classification is an integral part of deciding what properties it should be based upon.

*“A key property of classifications is that they can be nested within one another, creating a hierarchy. Thus, any group within a classification can be split in still greater detail. For example, objects that have been classified as chairs might be subdivided into stools, rocking chairs, recliners, etc. Similarly, listings in the "yellow pages" of a telephone directory are first classified by the product or service, and then within each of these categories they are further classified alphabetically. And there is no limit to the depth of a hierarchical classification. Most only have a few levels, but there are some that are quite deep”.* [Gary Olsen, 2003]

### **2.2.3. Semantic Relatedness**

Budanitsky and Hirst [Budanitsky, Alexander and Hirst Graeme, 2001], *“On the approach taken here, similarity and semantic distances in conceptual spaces are intimately connected. There is, however, a contrary view of the relationship between concepts and similarity”.*

The literature of Artificial Intelligence and computational linguistics is replete with approaches to measure semantic similarity (or more generally, semantic relatedness). Semantic relatedness

is a more general concept in the sense that it refers to whether two concepts are related (but not necessary) in some way.

Budanitsky and Hirst [Budanitsky, Alexander and Hirst Graeme, 2001] identified three categories of approaches to measuring semantic relatedness found in the literature: analytic approaches, statistical and machine learning approaches, and hybrid approaches.

Analytic approaches attempt to use physical properties of an ontological network of concepts such as WordNet online dictionary to estimate the similarity between concepts. Statistical and machine learning approaches attempt to use stored-up or learned statistics as the basis of judging similarity among concepts. Hybrid approaches combine both approaches to have some of the most effective means of assessing semantic similarity and semantic relatedness.

The need to determine the degree of semantic similarity, or, more generally, relatedness, between two lexically expressed concepts is a problem that pervades much of computational linguistics. Measures of similarity or relatedness are used in such applications as word sense disambiguation, determining discourse structure, text summarization and annotation, information extraction and retrieval, automatic indexing, lexical selection, and automatic correction of word errors in text.

In our experiment, like Budanitsky and Hirst [Budanitsky, Alexander and Hirst 2001], we have limited the scope of our search for semantic distance measures to those that make use of the

WordNet network as their knowledge source. This is due primarily to WordNet's availability and richness of supporting tools.

### 3. Methodology

First of all, WordNet gives the interface to list all the categories for a word.

For example, *close* and *deaths*, from WordNet, their presentations of hypernym are:

```
{orchid-yxzhao}wn deaths -hyphen
```

```
Synonyms/Hypernyms (Ordered by Frequency) of noun death
```

```
8 senses of death
```

```
Sense 6
```

```
death, dying, demise
```

```
6      => end, ending
```

```
5      => point, point in time
```

```
4      => time period, period, period of time
```

```
3      => fundamental quantity, fundamental measure
```

```
2      => measure, quantity, amount, quantum
```

```
1      => abstraction
```

In the presentation above, the list for word *deaths* has 8 senses, and sense6 has 6 levels in hierarchy.

```
{orchid-yxzhao}wn close -hyphen
```

```
Synonyms/Hypernyms (Ordered by Frequency) of noun close
```

```
3 senses of close
```

```
Sense 1
```

```
stopping point, finale, finis, finish, last, conclusion, close
```

```
6      => end, ending
```

```
5      => point, point in time
```

```
4      => time period, period, period of time
```

```
3      => fundamental quantity, fundamental measure
```

```
2      => measure, quantity, amount, quantum
```

```
1      => abstraction
```

In the presentation above, the list for word *close* has 3 senses, and sense1 has 6 levels in hierarchy.

Those two words have common categories and we can compare them by levels starting from roots. Here, we can have a constant ignoring *level*, this means we only care about the categories with their level higher than the ignoring *level*.

Then we will have a matrix showing the number of categories matched between each two words, and get the largest number *n* in that matrix. We will also have another constant *p* (weight percentage), and *n \* p* (*n* multiply *p*) is the cluster requirement that we only care the number of matched categories above the *n \* p*.

These two parameters are useful in reducing the processing size and also for the accuracy of the result.

For example, suppose *level* is 3, and *p* is 50%, for those two words *close* and *deaths*, we will ignore the category part as below:

3	=> fundamental quantity, fundamental measure
2	=> measure, quantity, amount, quantum
1	=> abstraction

We count their common categories starting from level 4 and *n* is 7. These two words are clustered because they have common categories. If we have more than two words, we will see the usage of *p* (See Page-33).



### 3.1. Definitions

#### Definition-3.1.

For any word  $w$  in a given text, this project is to collect all categories of  $w$  from all its senses of WordNet, and build a table with **category** and its **sense level** such as  $\{(w, (cate, level))\}$ .

Sense level is obviously in hypernym, in synonym and meronym all their sense levels are 1.

For example, for *close* and *deaths*, we will have a collection as below after we retrieve all their categories and their sense level in hypernym:

{ (death (end 6)(ending 6)(point 5)(point in time 5)(time period 4) (period 4) .....(abstraction 1))  
(close ((end 6)(ending 6)(point 5)(point in time 5)(time period 4) (period 4) .....(abstraction 1)) }

#### Definition-3.2.

We define  $\Delta C$  as the set of noun clusters,  $S(w)$  as set of category words of  $w$  of senses  $S$ .

If a cluster  $C$  (as one cluster of some  $w$ )  $\subseteq \{ S(w1) \cap S(w2) \}$

Then  $C \subseteq \Delta C$  and  $\{w1, w2\} \subseteq C$ .

For example, *close* and *deaths*, according to definition-1, we have  $S(close)$  and  $S(deaths)$ . And we can get  $S(close) \cap S(deaths)$ . If *close* and *deaths* are clustered in a cluster  $C$ , and  $C \subseteq S(close) \cap S(deaths)$ , then *close* and *deaths* are included in  $C$ , and  $C$  must be one cluster of  $\Delta C$ .

#### Definition-3.3

For clustering nouns, we are going to use the percentage constant  $p$ . If a cluster  $C \subseteq p * [ S(w1) \cap S(w2) ]$ , then  $C \subseteq \Delta C$  and  $\{w1, w2\} \subseteq C$ .

For example, *close* and *deaths*,  $C=\{\text{close, deaths}\}$  is not only a kind of  $C'=\{\text{end, ending}\}$ , but also a kind of  $C''=\{\text{point, point in time}\}$ .

#### Definition-3.4

Combination of word clusters with different semantic relations is the superset of the individual clusters with one semantic relation such as hypernym, synonym and meronym. This is expressed as below:

$$C(h) \subseteq C(h \cup s \cup m) \text{ or } C(s) \subseteq C(h \cup s \cup m) \text{ or } C(m) \subseteq C(h \cup s \cup m)$$

$h$  means hypernym,  $s$  means synonym and  $m$  means meronym

$C(h)$  means Cluster set of hypernym

$C(s)$  means Cluster set of synonym

$C(m)$  means Cluster set of meronym

The sign  $\cup$  means union.

#### Definition-3.5

Suppose,  $C1(h \cup s \cup m) = \{w1, w2, \dots, wi\}$ ,  $C2(h \cup s \cup m) = \{w2, w3, \dots, wj\}$

Then we can combine  $C1$  and  $C2$  as:  $C3(h \cup s \cup m) = \{w1, w2, w3, \dots, wi, wj\}$ , the combined words will be unique.

And also we have the result as definition-3.4.

The sign  $\cup$  means union.

$C1, C2, C3$  are Clusters.

## 3.2. Algorithm

### 3.2.1. Input

Input: *Tagged File Name* (eg. FT923-6455.tagged)

Input: Percentage constant  $p$  (start from 0 to 99)

Input : Ignoring *level* (start from 0)

### 3.2.2. Working flow

1. **foreach** noun
2. get all items of all senses with sense level from WordNet
3. put noun in *list*
3. **foreach** noun1 in list
7. **foreach** noun2 in list
8. compute similarity (verify their matches of the result of step 2) of noun1 and noun2 and create matrix
9. **foreach** noun1, noun2 in matrix
10. **if** similarity > threshold (user's given)
11. assign noun1, noun2 to a cluster

Now we can demonstrate the algorithm with an example for two words as *deaths* and *close*.

First of all we can easily get the information from WordNet stored in the structure as

{word (level category) (level category) ...}, and will get the noun list as {"deaths" "close"}.

Now we compare two words *deaths* and *close* in the noun list with their retrieved information, and create a matrix as below:

	Close
Death	10

The above means *deaths* has 10 matched category items with *close*.

And 10 is the largest number in this matrix, then we can compute out the threshold =  $10 * 50\%$   
=5.

Now the final work is to pick up all those words with number of category matched greater than or equal to 5 from the matrix and put those words into a cluster, thus we can get the cluser  $\{(death, close)\}$ .

In this project, we defined 5 tables to store the intermediate data during the whole processing. The following will describe their usage and all their procedures.

### 3.2.2. Tables

#### 1. Table-1.Word index table

Attributes : ID, Word

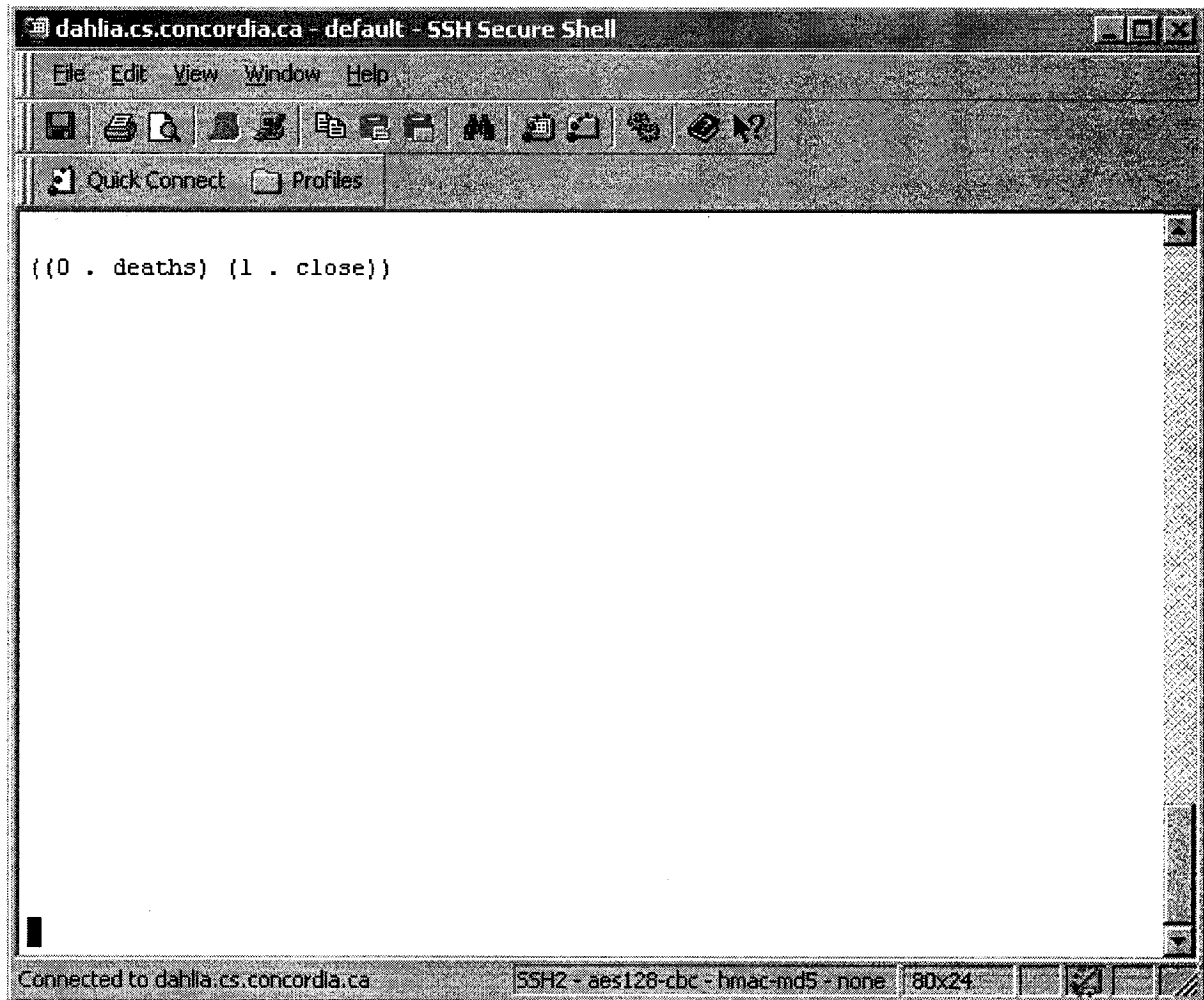
Note : Distinct Words and Category Retrieval successfully through WordNet.  
ID start from 0, increasing by 1

Usage : Use to create matrix from category table.

e.g.  $\{ (0 \text{ deaths}) (1 \text{ close}) \}$ . Any word inside this table means it has retrieved categories from WordNet. This table will be used to create matrix.

Figure-1 showed the performance in this processing step of these two words.

**Figure-1 .Noun and Index List**



## **2. Table-2. Category table**

Attributes : Word, Category, Type, Level

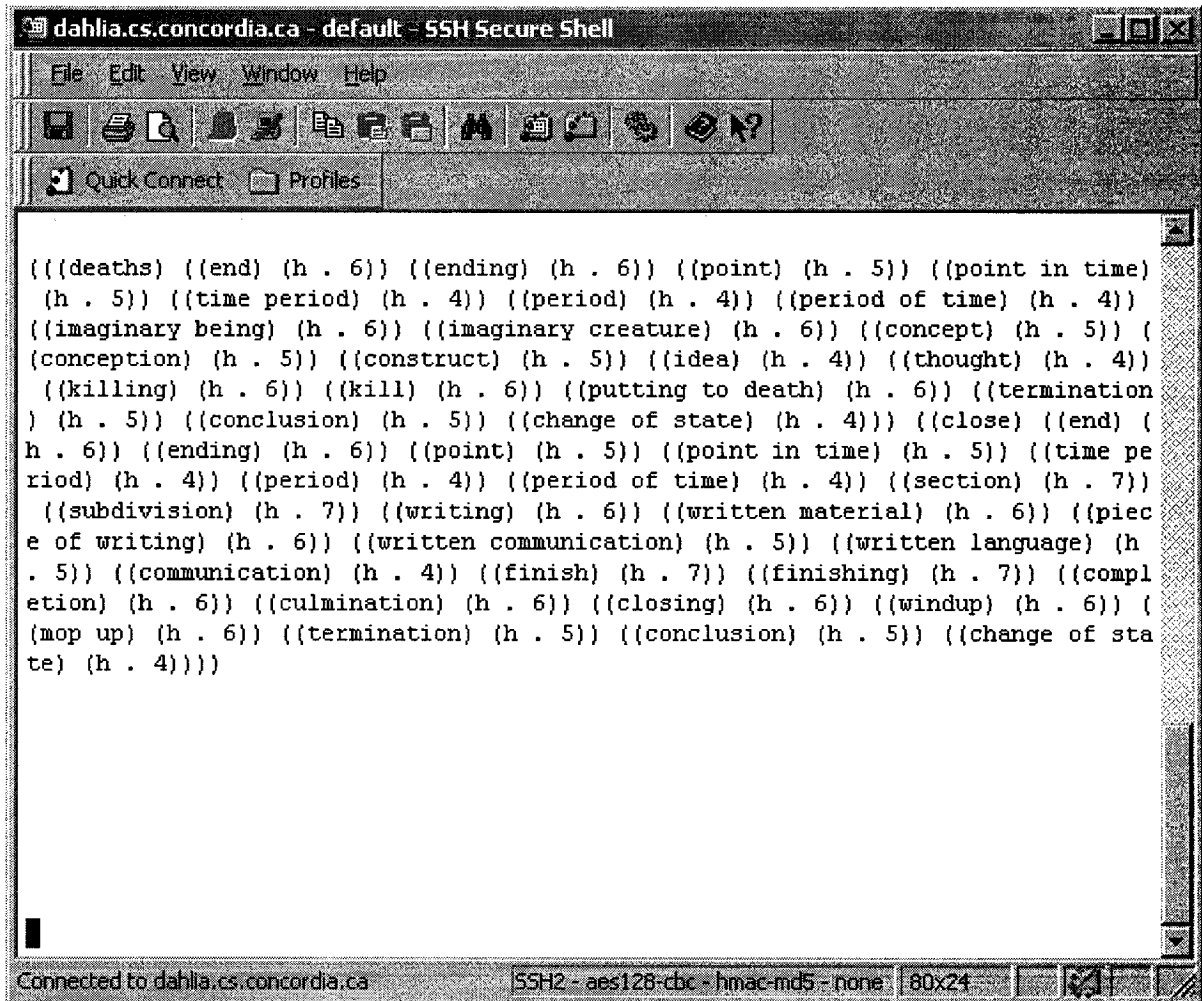
Note : Main table to hold detailed category information with category level being greater than the ignoring *level* for each word. Type is used to express the semantic relation. For example, h means hypernym.

Usage : Used to generate matrix.

e.g. ( ( (deaths) ( (end) (h . 6) ) ( (ending) (h . 6) ) ... )  
 ( (close) ( (end) (h . 6) ) ( (ending) (h . 6) ) ... )

This table will be used to compare their semantic relatedness. See Figure-2

**Figure-2. Category and Level**



In Figure-2, it showed all the categories with their pair of type and level for *deaths* and *close*.

### 3. Table-3. Matrix table

Attributes : Word, WordComparedWith, NumOfCategoryMatched

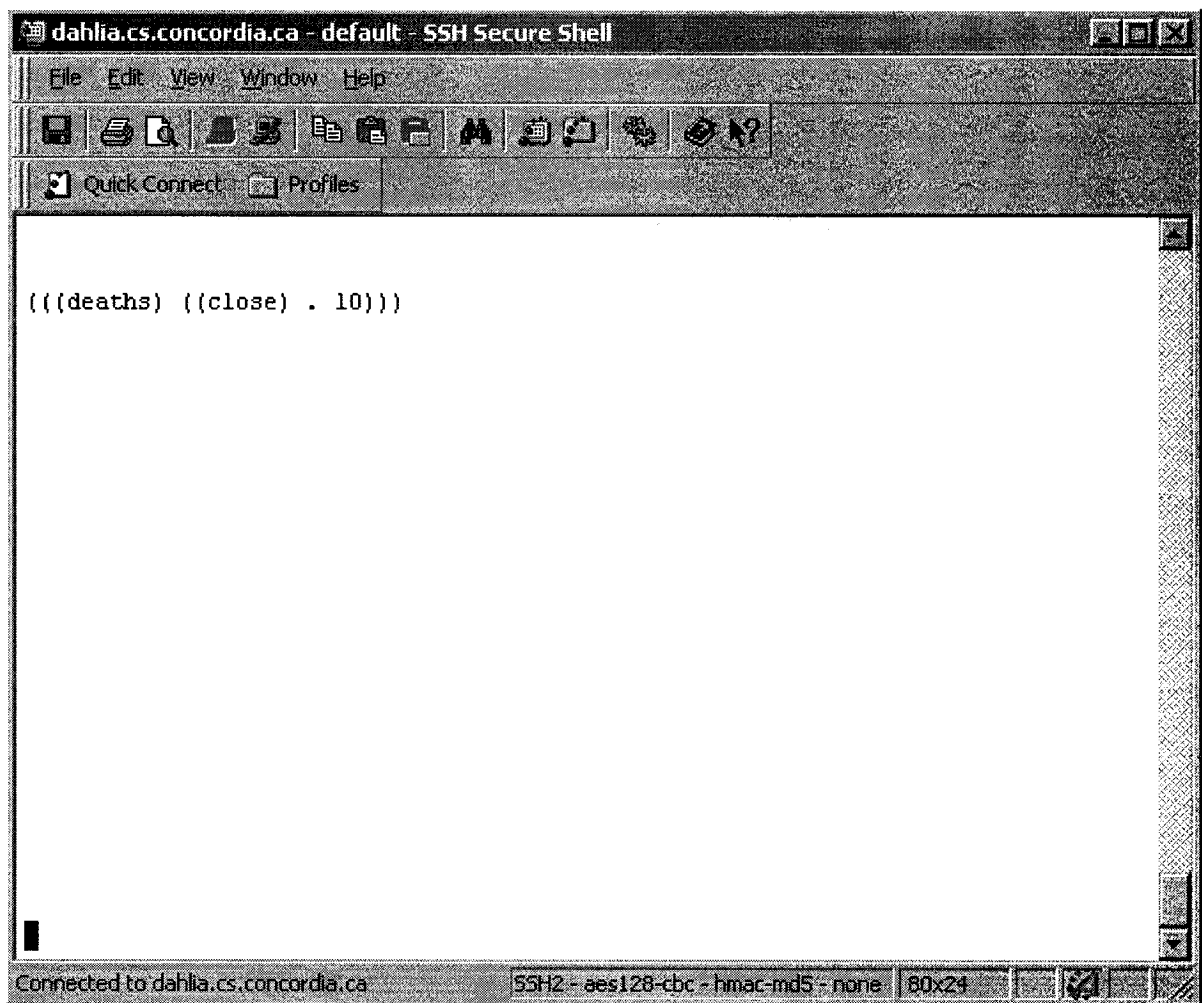
Notes : the matrix is going to be in one direction  $w_1w_2$ ,  $w_1w_3$ ,  $w_n-1w_n$ .

Usage : Used to analyse the result and to generate clusters.

e.g. { (deaths close 10) ... }

This table will be used to hold the information before creating the matrix. See Figure-3.

**Figure-3. Nouns and Number of Common Categories**



In Figure-3, it showed that *deaths* and *close* have 10 common categories retrieved through WordNet.

#### **4. Table-4. Rough Clusters table**

Attributes :  $Word_n, Word_{n+1}$

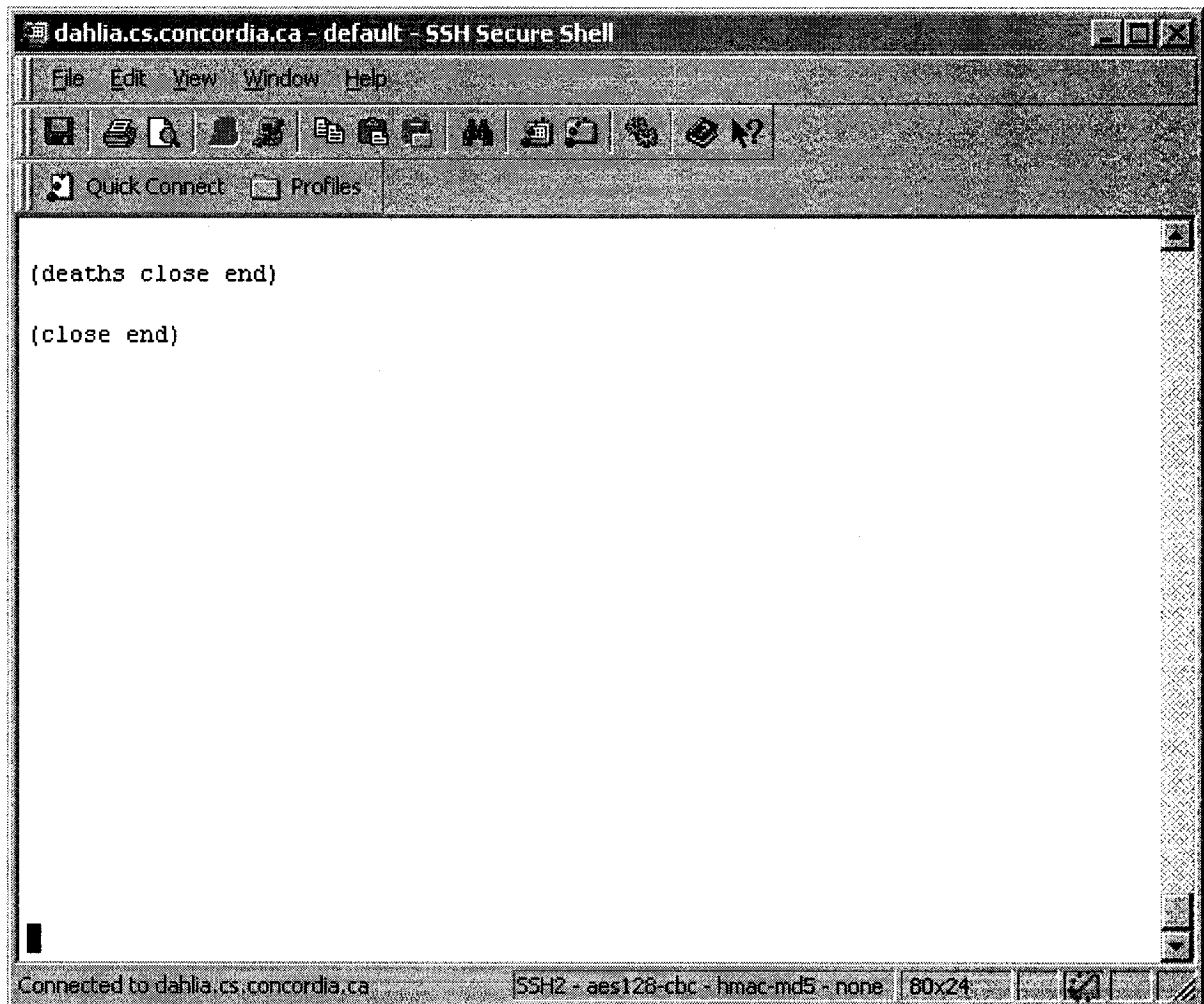
Usage : Used for cluster optimization

e.g. { (deaths close...)

This table will be used to reduce the redundant clusters. For example, if we have two clusters as (deaths close end) (close end), the final cluster will be one cluster as (deaths close end).

Figure-4 showed all the clusters stored in table-4.

**Figure-4. Clusters**



## 5. Table-5. Optimized Clusters table

Attributes :  $Word_n$ ,  $Word_{n+1}$

Usage : used to remove the redundant clusters with their elements having chained words.

e.g. { (deaths close end...)} should be optimized cluster.

This table will be used to store the final clusters.



### 3.2.3. Process stages

There are four stages during the processing of a document. They are Noun retrieval, syntactical extraction, matrix generation and clusters generation.

#### 1. Stage1: Noun retrieval from tagged text.

**Function1:** Retrieve nouns from tagged text.

Input: file name of tagged text

Output: all the nouns of input text

This is to extract fragments of the corpus that match a syntactic pattern. Those words matching the pattern, such as Word/NN, Word/NNP and Word/NNS, will be extracted. The fragments are output to a file for future use. Duplicate nouns will be ignored.

Figure-5 gives an example of how a tagged file looks like (Input: FT923-6455.tagged):

This stage will only pick up sequences of nouns inside the text's body (quoted with <TEXT> and </TEXT>) such as: "Mr Justin Balcombe", "insurance", "adjuster", "losses"..., etc. and will pass them to stage 2.

In the tagged file's text body, all those words suffixed with /NN or /NNS or /NNP are nouns, and we will group all those words suffixed with NNP continuously as noun phrase.

Figure-5

**Figure-05 Tagged Text Format**

.....  
.....  
Mr/NNP Justin/NNP Balcombe/NNP ,/, of/IN UK-based/JJ insurance/NN  
adjuster/NN Balcombe/NNP Group/NNP ,/, said/VBD  
total/JJ losses/NNS could/MD exceed/VB Dollars/NNPS 15bn/CD if/IN  
business/NN interruption/NN claims/NNS were/VBD  
taken/VBN into/IN account/VBD That/DT compares/VBZ with/IN the/DT  
Dollars/NNPS 4bn-Dollars/NNS 6n/CD (Pounds/NNS  
2.1bn-Pounds/NNS 3.1bn)/CD of/IN insurance/NN industry/NN losses/NNS  
caused/VBN by/IN the/DT last/JJ big/JJ US/PRP  
hurricane/NN ,/, Hugo/NNP ,/, which/WDT hit/VBD South/NNP Carolina/NNP in/IN  
1989./CD  
The/DT brunt/NN of/IN the/DT losses/NNS are/VBP likely/JJ to/TO be/VB  
concentrated/VBN among/IN US/PRP insurers/VBP  
industry/NN analysts/NNS said/VBD yesterday./JJ  
.....  
.....

**2. Stage2: Creating category list.**

**Function2:** Retrieve all category information in WordNet for the words extracted from tagged text.

Input: nouns from tagged text

Output: create a temporary file to store the retrieved information from WordNet interface for each noun.

**Function3:** Generate pair of information with category and sense level

Category information will be processed in each sense paragraph, and reverse the category level to make sure the top level is the leaf and bottom is the root (especially for hypernym).

Same category and level will not be accepted to insert into the table.

Input: loading temporary file

Output: prepare distinct pair of information with category and its sense level before inserting into Table-1

**Function4:** Filter hypernym level.

All the categories with the sense level less than user's threshold won't be inserted into Table-1 (only for hypernym).

Input: prepared pair of information from Function-3

Output: filter each pair by checking its sense level against the user's threshold value and prevent that word from being inserted into Table-1

**Function5:** Insert successful nouns into Table-2.

During the above processes for each noun, meanwhile create Table-1. All the noun words with their retrieved information from WordNet will be inserted into Table-1. This will be used in making the matrix.

Input: select all the distinct nouns in Table-1

Output: insert them into Table-2.

### **3. Stage3: Making the matrix**

**Function6:** Calculate NumOfCategoryMatched

Using each two words from Table-2 to query the NumOfCategoryMatched from Table-1 and insert them into the matrix of Table-3.

Input: select each pair of nouns from Table-2 to match their information in Table-1.

Output: insert the matched result (include those nouns and the number of matched category) into Table-3

**Function7:** Calculate MaxNumOfCategoryMatched

This will be used to make clusters in Stage3.

The NumOfCategoryMatched will be ignored for calculating the MaxNumOfCategoryMatched when the two words are in the following cases: (1). Plural and singular. (2). Upper case and lower case. (3). (1)+(2).

Here, for example, we will not distinguish between *Interest* and *interest*, *interests* and *interest*.

Input: select each pair of nouns from Table-2 to match their information in Table-1.

Output: insert the matched result (include those nouns and the number of matched category) into Table-3

**4. Stage4: Making sub-clusters and final combination clusters**

**Function8:** Making rough sub clusters

Compare each two nouns ( $wd_n$ ,  $wd_{n+1}$ ) from Table-3 to retrieve those records with NumOfCategoryMatched larger than the user's request weight ( $= p$  (weight percentage) \* MaxNumOfCategoryMatched / 100), and then insert the records into Table-4.

Input: select all the records from Table-3 including pairs of nouns and their NumOfCategoryMatched

Output: insert those records with NumOfCategoryMatched  $\geq (p * n)$  into Table-4, and then the rough clusters will be generated in the format:

{ ( $wd_1$  ( $wd_2$  NumOfCategoryMatched1) ( $wd_2$  NumOfCategoryMatched2)...) ( $wd_2$  ( $wd_3$  NumOfCategoryMatched3) ( $wd_4$  NumOfCategoryMatched4)...)...}

**Function9:** Making optimized sub clusters.

Clusters such as (wdn, wdn+1....) (wdn+1, wdn+2....) will be optimized as (wdn, wdn+1, wdn+2 ....). This function will be processed recursively until all the clusters have been completely optimized.

Input: select all the records from Table-4

Output: appending those clusters with first noun existing inside another cluster. This function will remove redundant clusters.

**Function10:** Making optimized sub clusters with cluster length in descending orders.

**Function11:** Making final clusters

Combine all three optimized sub clusters as final clusters.

Process Function9 and Function10 to get optimized final clusters stored in Table-5.

### **3.3. Scheme Code**

#### **3.3.1. Gambit**

This project was developed using the Gambit Scheme language under the Unix system.

The Gambit system (including the Gambit-C version) is Copyright (C) 1994-1998 by Marc Feeley (Dépt. d'informatique et r.o. Université de Montréal). The Gambit programming system is a full implementation of the Scheme Language which conforms to the R4RS and IEEE Scheme standards. It consists of two programs: 'gsi', the Gambit Scheme interpreter, and 'gsc', the Gambit Scheme compiler.

*“Gambit-C is a version of the Gambit system in which the compiler generates portable C code, making the whole Gambit-C system and the programs compiled with it easily portable to many*

*computer architectures for which a C compiler is available. Several extensions to the standards are provided. It includes different versions for kinds of platforms, such as UNIX, PC and Macintosh.”* (<http://www.iro.umontreal.ca/~gambit>)

## **1. Gambit system Unix interface**

To run the Gambit Scheme interpreter:

```
gsi [-:RUNTIMEOPTION,...] [-f] [-i] [-e EXPRESSIONS] [FILE...]
```

To run the Gambit Scheme compiler:

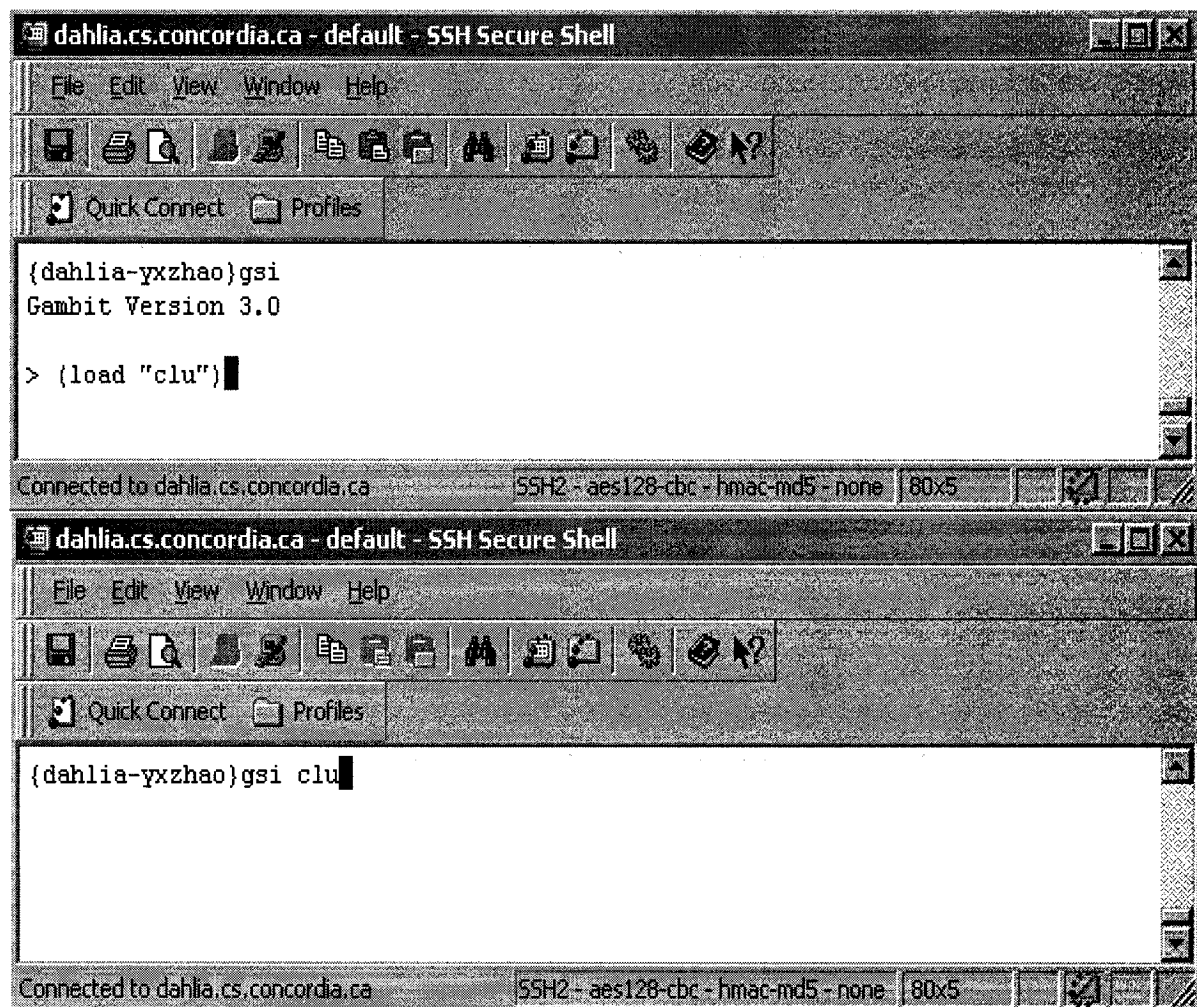
```
gsc [-:RUNTIMEOPTION,...] [-f] [-i] [-e EXPRESSIONS] [-prelude EXPRESSIONS]  
    [-postlude EXPRESSIONS] [-verbose] [-report] [-expansion] [-gvm] [-debug]  
    [-o OUTPUT] [-c] [-dynamic] [-flat] [-l BASE] [FILE...]
```

There are two ways to use the interpreter or compiler. First way is without scheme program name. It is useful to debug the scheme code. Once an error occurred, all the information stays in the memory. The second way is with the scheme program name. Once an error occurred, the cursor returns to command line and without debugging. See Figure-6.

An error is not signaled if the file does not exist.

The interface of Gambit-C Scheme system version 3.0 is shown as below (Figure-6):

**Figure-6** (Gambit system interface for Unix system)



### 3.3.2. Scheme Code

In this project, many looping performances enhancements have been used either in retrieval progress or in optimising progress. In order to improve the efficiency of performance, and the processing speed, all the looping performances are processed in a tail recursion. For the Scheme language, we mainly refer to the book of Brian Harvey and Matthew Wright [Brian Harvey and Matthew Wright, 1994].

In Appendix-A, there are two procedures are attached to demo this project.

For example,

```
(define create_matrix  
  (lambda ()  
    (if (ok)  
        (return ok)  
        (create_matrix)))  
  )  
)
```

### 3.3.3. Performance Interface

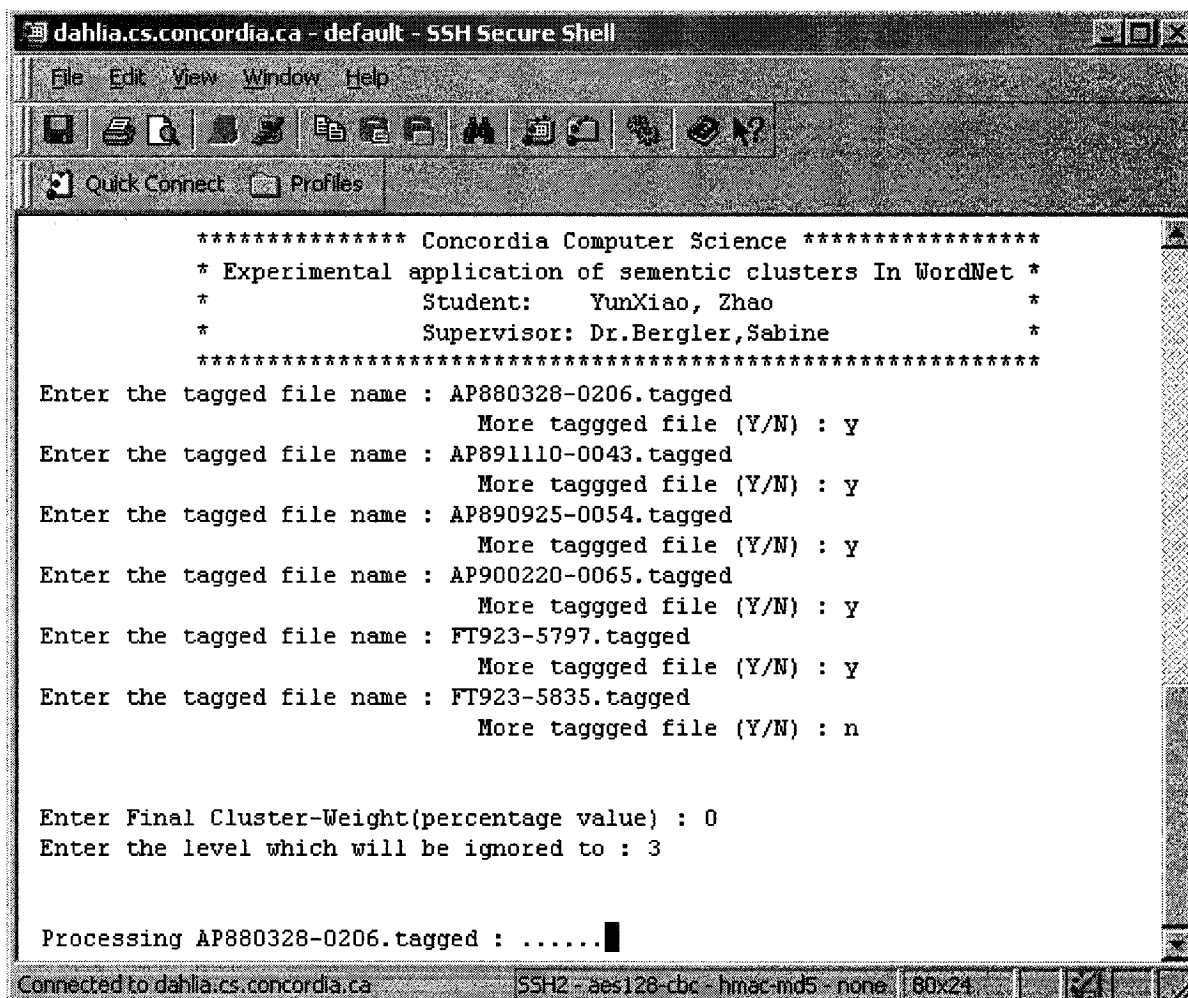
As the performance interface shown below (Figure-7), we process the tagged document with **two required parameters**. One is the cluster-weight (as introduced in Definition-3.3) and the other one is the level (as introduced in Definition-3.1).

**Three semantic relations** will be processed one by one in the order as: hypernym, synonym and meronym, and **three processing steps** are displayed during the processing as:

- Creating category list
- Generating matrix
- Creating the clusters.



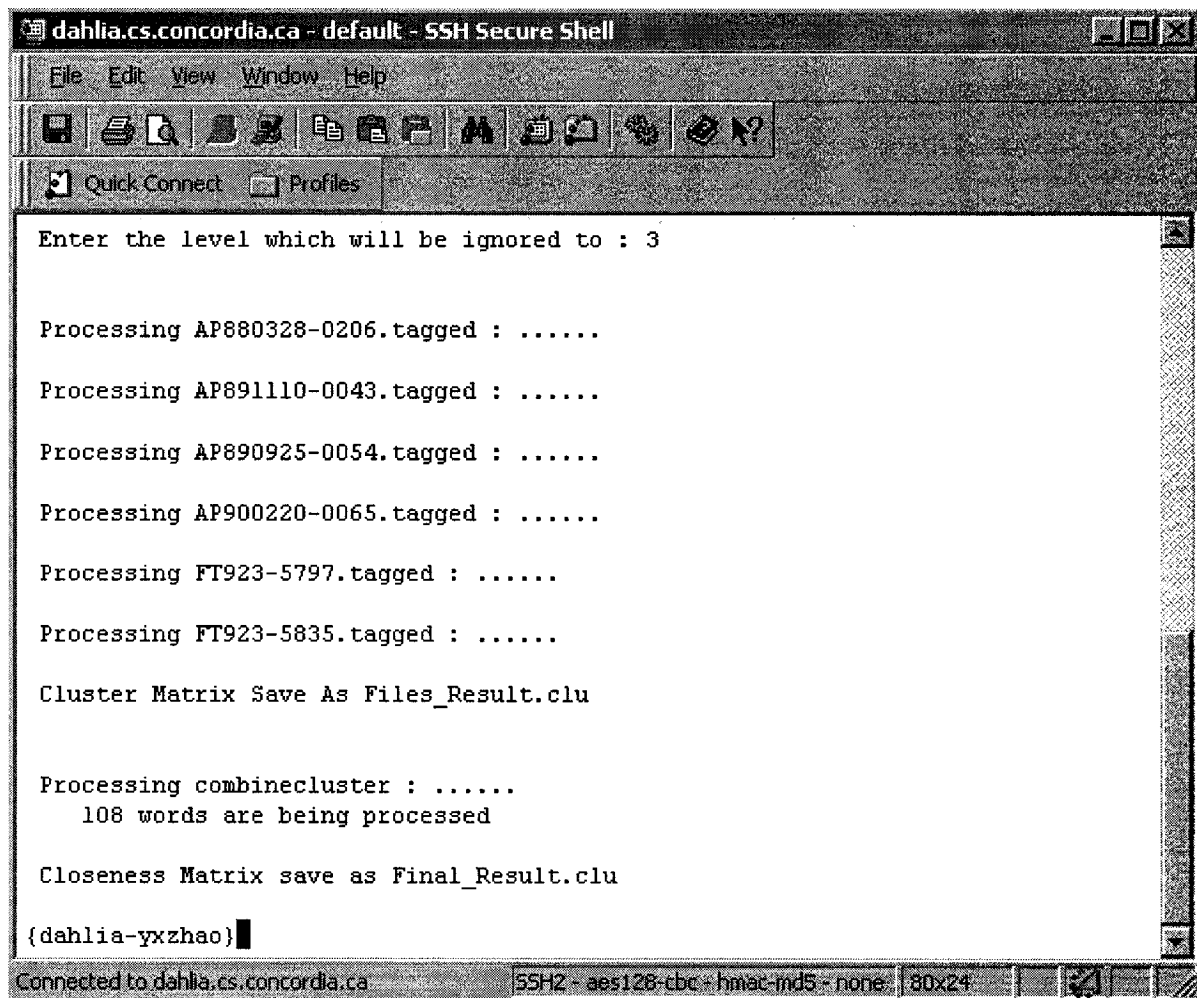
Figure-7 Processing Interface



In Figure-7, we allow user to enter multiple tagged files to process and require two parameters.

Once the processing started, there are generally two processing steps: one step is processing single tagged file and another step is processing the combination of the results of each single processed tagged file. The results of these two steps will be saved into different files (shown in Figure-8). If we give one document, we will only have the first processing step.

**Figure-8** End Processing Interface



The image shows a terminal window titled "dahlia.cs.concordia.ca - default - SSH Secure Shell". The window has a menu bar with "File", "Edit", "View", "Window", and "Help". Below the menu bar is a toolbar with various icons. The main area of the terminal displays the following text:

```
Enter the level which will be ignored to : 3

Processing AP880328-0206.tagged : .....
Processing AP891110-0043.tagged : .....
Processing AP890925-0054.tagged : .....
Processing AP900220-0065.tagged : .....
Processing FT923-5797.tagged : .....
Processing FT923-5835.tagged : .....

Cluster Matrix Save As Files_Result.clu

Processing combinecluster : .....
    108 words are being processed

Closeness Matrix save as Final_Result.clu

{dahlia-yxzhao}
```

At the bottom of the terminal window, a status bar shows "Connected to dahlia.cs.concordia.ca", "SSH2 - aes128-cbc - hmac-md5 - none", and "80x24".

## 4. Results and Evaluations

For experiments and evaluations of this project's purpose, here we took 6 tagged documents and processed them with different cluster weight percentage as 0%, 20% and 40% and the category ignoring level was kept the same at 3.

These 6 documents are 2 of economy, 3 of entertainment, and 1 of natural disaster.

First of all, we will process the noun clustering of each document, clustering weight percentage from each category matrix of Hypernym, Synonym and Meronym is 0%. This means that we will cluster all those words in the matrix.

Secondly we are going to combine those clusters that each of them is the largest one in that of document, and process noun clustering of the combined cluster as a document, the performance is the same as step 1.

Finally, based on the result after step 2, we will fetch three groups of clusters with clustering weight percentage 0%, 20% and 40%, and on this step will generate the final analysis report and summarize analysis report, also we are going to evaluate these reports.

#### **4.1. Results and Analysis of Single Document Clustering**

For each processed tagged file, four matrixs will be generated as the following sample format of the first tagged file. There is a matrix for hypernym, a matrix for synonym, a matrix for meronym and the matrix for combination of hypernym, synonym and meronym.

All the clusters of the six processed tagged files have been overlapped based on rough clusters of hypernym, synonym, hypernym and their combination generated from their matrixs.

The contents of all the tagged files are appended in Appendix B.

### 4.1.1. Sample-1 (Entertainment-AP880328-0206)

#### (1). Matrix of Hypernym, Synonym and Meronym

Notes:

Wd1 --> HOLLYWOOD

Wd2 --> wealth

Wd3 --> years

Wd4 --> Academy

Wd5 --> actress

Wd6 --> actor

Wd7 --> director

Wd8 --> film

Wd9 --> Virginia

Wd10 --> studio

Wd11 --> performer

Wd12 --> award

Wd13 --> Sunshine

Wd14 --> team

Wd15 --> Oscars

Wd16 --> Sierra

Wd17 --> golf

Wd18 --> statuette

Wd19 --> year

Wd20 --> people

Wd21 --> record

Wd22 --> reading

Wd23 --> winner

\*\*\*\*\* Matrix of Nouns for hypernym \*\*\*\*\* AP880328-0206.tagged

===== ( hyphen level ignored to: 3 ) =====

	Wd2	Wd3	Wd4	Wd5	Wd6	Wd7	Wd8	Wd9	Wd10	Wd11	Wd12	Wd13	Wd14	Wd15	Wd16	Wd17	Wd18	Wd19	Wd20	Wd21	Wd22	Wd23
Wd1 -->	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Wd2 -->		2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Wd3 -->			0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
Wd4 -->				0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Wd5 -->					3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Wd6 -->						0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Wd7 -->							0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Wd8 -->								0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
Wd9 -->									0	0	0	0	0	0	0	0	0	0	0	0	6	0
Wd10 -->										0	0	0	0	0	0	0	0	0	0	0	0	0
Wd11 -->											0	0	0	0	0	0	0	0	0	0	0	0
Wd12 -->												0	0	7	0	0	0	0	0	5	5	0
Wd13 -->													0	0	0	0	0	0	0	0	0	0
Wd14 -->														0	0	0	0	0	0	0	0	0
Wd15 -->															0	0	0	0	0	5	5	0
Wd16 -->																0	0	0	0	0	0	0
Wd17 -->																	0	0	0	0	0	0
Wd18 -->																		0	0	0	1	0
Wd19 -->																			0	0	0	0
Wd20 -->																				0	0	0
Wd21 -->																					5	0
Wd22 -->																						0

(Note:  $n=7$ ,  $n \cdot p=7 \cdot 0=0$  See Page-13 definition of  $n$  and  $p$ )

In the matrix above, it showed the comparison results between each two nouns with the weight percentage  $p$  being 0% and the ignoring *level* being 3, the highlighted number 7 is  $n$ .

So, from this matrix, we clustered those words with their number of common categories being non-zero and greater than or equal to  $n * p$ .

**(2). Clusters of Hypernym (four clusters after being overlapped)**

```
(film record award Oscars reading Virginia statuette)
(wealth years year)
(actress actor performer)
(Academy studio)
```

**(3). Clusters of Synonym (only two clusters)**

```
(years year)
(actor winner)
```

**(4). Clusters of Meronym (only one cluster)**

```
(years year)
```

**(5). Combined Clusters (four clusters after being combined and overlapped)**

```
(film record award Oscars reading Virginia statuette)
(actress actor winner performer)
(wealth years year)
(Academy studio)
```

**(6). Analysis of result**

All the clusters above are correct except cluster (film record award Oscars reading Virginia statuette).

Inside this cluster, *Viginia* and *reading* are clustered because WordNet is not case sensitive. *Reading* and *reading* in WordNet have the same description. And Reading has the meaning of region, geographical area, city, etc. We can't skip this sense of *Reading* because it is not distinguished from other similar senses.

#### **4.1.2. Sample-2 (Entertainment-AP891110-0043)**

##### **(1). Clusters of Hypernym (only one cluster after being overlapped)**

(statuette images copyright statue part trophy shock protection cap decision industry film service trademark star finish award mark fashion symbol world manufacturer ruling judge Jan. date century Academy academy court)

##### **(2). Clusters of Synonym (five clusters after being overlapped)**

(world manufacturer industry domain part protection ceremony service film)  
(trademark mark award)  
(ruling decision)  
(setback finish)  
(Academy academy)

##### **(3). Combined Clusters (only one cluster after being combined and overlapped)**

(statuette images copyright statue part trophy shock protection cap decision industry film service trademark star finish award mark fashion symbol world manufacturer ruling judge Jan. date century Academy academy court)

##### **(4). Analysis of result**

All the clusters above are correct. Nouns *finish* and *symbol* are clustered because they have the common category as *basic cognitive process* in sense level 4 of hypenym. They will not be clustered with the increasing of sense level or clustering weight percentage. They have no more common categories in upper levels.

### 4.1.3. Sample-3 (Natural Disaster-AP890925-0054)

#### (1). Clusters of Hypernym (two clusters after being overlapped)

(word lives respect close evacuation damage injury time news area brunt mph  
number fury billions thousands tens Tens lot)

(storm hurricane)

#### (2). Clusters of Synonym (four clusters after being overlapped)

(billions thousands tens Tens lot)

(lives victims time mph)

(word news)

(damage injury)

#### (3). Combined Clusters (two clusters after being combined and overlapped)

(word lives victims respect close evacuation damage injury time news area  
brunt mph number fury billions thousands tens Tens lot)

(storm hurricane)

#### (5). Analysis of result

Noun *word* and *lives* are clustered because they have just one common category item *communication* in level 4 (they won't be clustered with cluster weight percentage increasing to 20%, 40%). And they have no more leaves up to level 4. If we can uncluster this kind of nouns, then the final results (see Page 46/47) won't show us that this document is close to the next one of Sample-4.

### 4.1.4. Sample-4 (Entertainment-AP900220-0065)

#### (1). Clusters of Hypernym (only one cluster after being combined and overlapped)

(cap sweatshirt reception lecture session film picture characters movie  
ones)



**(2). Clusters of Synonym** (only one cluster after being overlapped)

(movie picture film lot)

**(3). Clusters of Meronym** (only one cluster after being overlapped)

(movie picture film)

**(3). Combined Clusters**

(cap sweatshirt reception lecture session film picture characters movie ones lot)

**(4). Analysis of result**

All the clusters above are correct

**4.1.5. Sample-5 (Economy-FT923-5797.tagged)**

**(1). Clusters of Hypernym** (three clusters after being overlapped)

(SQUADS troops distribution government farming agriculture state homes quarter cent department Florida people cotton water alligator oil emergency secretary officials week night thousands count estimates deliveries decline fury)

(hurricane storm rainstorm)

(rebuilding destruction)

**(2). Clusters of Synonym** (four clusters after being overlapped)

(week night quarter cent)

(state agriculture farming department)

(SQUADS troops homes)

(distribution government)

**(3). Clusters of Meronym** (only one cluster)

(farming agriculture)

**(4). Combined Clusters** (three clusters after being combined and overlapped)

(SQUADS troops distribution government farming agriculture state homes  
quarter cent department Florida people cotton water alligator oil emergency  
secretary officials week night thousands count estimates deliveries decline  
fury)  
(hurricane storm rainstorm)  
(rebuilding destruction)

**(5). Analysis of result**

All the clusters above are correct

**4.1.6. Sample-6 (Economy-FT923-5835.tagged)**

**(1). Clusters of Hypernym** (four clusters after being overlapped)

(insurer company meeting cent insurance information share level threshold  
branch area time respect reach)  
(chairman manager)  
(basis property)  
(losses loss)

**(2). Clusters of Synonym** (three clusters after being overlapped)

(reach area level branch)  
(meeting company information)  
(losses loss time)

**(3). Combined Clusters** (four clusters after being combined and overlapped)

(insurer company meeting cent insurance information share level threshold  
branch area time losses loss respect reach)  
(chairman manager)  
(basis property)

**(4). Analysis of result**

As shown above, *area* and *property* should be clustered in the common category *region* because they have the common category item *region* in category level 3, but unfortunately they are ignored by our entered threshold value.

#### 4.1.7. Combination Cluster

The following are the processing results of the combination of the six largest clusters. It means that we combine the six largest clusters from each processed tagged file, continually process the combined clusters with distinct nouns and get the following results.

Actually, the combined clusters are used to compare each two documents later. We have analysed the processed results of the six files. So the following just gives the processed results.

##### (1). Clusters of Hypernym (only one cluster after being overlapped)

(film record award Oscars reading images copyright trophy shock cap  
trademark star mark symbol century word Virginia court lot billions  
thousands tens Tens ones quarter cent count state government department  
decision industry ruling fury characters night estimates Florida company  
branch lives respect close news number reception lecture session picture  
homes emergency secretary deliveries insurance information share level  
threshold statue part finish fashion service evacuation injury damage  
meeting statuette time mph world manufacturer farming agriculture  
distribution people cotton protection sweatshirt oil water alligator judge  
SQUADS troops insurer week movie Academy academy brunt reach date area Jan.  
officials decline)

##### (2). Clusters of Synonym (six clusters after being overlapped)

(film part protection service insurance cap share deliveries world industry  
manufacturer lives time agriculture people farming state department lot  
characters award mark trademark number branch brunt troops secretary symbol  
fashion mph session Oscars trophy respect date quarter week night homes  
statuette ones SQUADS company decision ruling estimates court meeting word  
news lecture area level reach Academy academy threshold distribution  
government water billions century thousands tens Tens cent decline  
information picture reading images movie)

(record copyright count)  
(Virginia Florida)  
(finish close)  
(judge officials)  
(damage injury)

### (3). Clusters of Meronym (two clusters after being overlapped)

(film picture movie)  
(farming agriculture)

### (4). Combined Clusters (only one cluster after being combined and overlapped)

(film part protection service insurance cap share deliveries world industry  
manufacturer lives time agriculture people farming state department lot  
characters award mark trademark number branch brunt troops secretary symbol  
fashion mph session Oscars trophy respect date quarter week night homes  
statuette ones SQUADS company decision ruling estimates court meeting word  
news lecture area level reach Academy academy threshold distribution  
government water billions century thousands tens Tens cent decline  
information picture reading images movie record copyright count shock star  
Virginia Florida fury close finish reception emergency statue evacuation  
injury damage cotton sweatshirt oil alligator judge officials insurer Jan.)

Through the results and the analysis of the six processed documents above, we realized that we can roughly tell the document's brief meaning from its largest cluster. And the more nouns are clustered, the more clear about the document's meaning from its largest cluster, but we need to only pick up those closer words to cluster. This will depends on the lexica structure.

Actually, before presenting those results above with ignoring level 3, we have tried 2 or 4 or 5.

In WordNet, all those items with category level less or equal to 2 are very common such as:

3       =>artifice, arteface  
2       =>object, physical object  
1       =>entity  
2       =>whole, whole thing, unit  
1       =>entity

or:

2       =>social group  
1       =>group, grouping

We will almost cluster all the nouns if we choose the items at least include category levels including from 3.

On the other hand, we also tried category level from 4 or 5, then we will get small clusters and many nouns will not be clustered.

## **4.2. Documents Closeness Analysis**

As we have assumed that we got the noun clusters correctly for single document, we continue to use the same algorithm to combine the nouns from the largest clusters of any two documents.

If we treat the clusters of a document as a super cluster, then we will have the clusters for each of the processed documents (one for each) using the cluster algorithm described above, and then we can continue clustering the super clusters. So, each cluster is actually a cluster of documents and an associated cluster of words, thus it is a document-word-co-cluster. Finally we will have the result of comparing each two documents.

We are trying to discover the closeness of each two documents by increasing the clustering weight percentage.

The following 3 pairs of analysis reports are automatically generated by this project with clustering weight percentage equal to 0% initially, increasing to 20% and to 40%.

\*\*\*\*\* Comparison of 6 tagged files \*\*\*\*\*

Percentage Weight (p)	Tagged File	Length of Largest Cluster	Words in Largest Cluster of Combination
0 %	AP880328-0206.tagged	7	7
	AP891110-0043.tagged	30	27
	AP890925-0054.tagged	19	19
	AP900220-0065.tagged	10	8
	FT923-5797.tagged	28	26
	FT923-5835.tagged	14	10
	Combination of above	97	97

From above, we fetch each two value in last column to derivate the following closeness matrix report.  
The formula of closeness value is:  $\text{Max}(\text{Value1}, \text{Value2}) / (\text{Value1} + \text{Value2})$ .

\*\*\*\*\* Closeness Matrix of 6 tagged files \*\*\*\*\*

Tagged Files	F 2	F 3	F 4	F 5	F 6
AP880328-0206.tagged [F 1]	27/34	19/26	8/15	26/33	10/17
AP891110-0043.tagged [F 2]		27/46	27/35	27/53	27/37
AP890925-0054.tagged [F 3]			19/27	26/45	19/29
AP900220-0065.tagged [F 4]				13/17	5/9
FT923-5797.tagged [F 5]					13/18
FT923-5835.tagged [F 6]					

\*\*\*\*\* Comparison of 6 tagged files \*\*\*\*\*

Percentage Weight(p)	Tagged File	Length of Largest Cluster	Words in Largest Cluster of Combination
20 %	AP880328-0206.tagged	7	7
	AP891110-0043.tagged	30	24
	AP890925-0054.tagged	19	17
	AP900220-0065.tagged	10	8
	FT923-5797.tagged	28	21
	FT923-5835.tagged	14	9
	Combination of above	86	86

From above, we fetch each two value in last column to derivate the following closeness matrix report.  
The formula of closeness value is:  $\text{Max}(\text{Value1}, \text{Value2}) / (\text{Value1} + \text{Value2})$ .

\*\*\*\*\* Closeness Matrix of 6 tagged files \*\*\*\*\*

Tagged Files	F 2	F 3	F 4	F 5	F 6
AP880328-0206.tagged [F 1]	24/31	17/24	8/15	3/4	9/16
AP891110-0043.tagged [F 2]		24/41	3/4	8/15	8/11
AP890925-0054.tagged [F 3]			17/25	21/38	17/26
AP900220-0065.tagged [F 4]				21/29	9/17
FT923-5797.tagged [F 5]					7/10
FT923-5835.tagged [F 6]					

\*\*\*\*\* Comparison of 6 tagged files \*\*\*\*\*

Percentage Weight (p)	Tagged File	Length of Largest Cluster	Words in Largest Cluster of Combination
40 %	AP880328-0206.tagged	7	6
	AP891110-0043.tagged	30	22
	AP890925-0054.tagged	19	15
	AP900220-0065.tagged	10	6
	FT923-5797.tagged	28	20
	FT923-5835.tagged	14	9
	Combination of above	78	78

From above, we fetch each two value in last column to derivate the following closeness matrix report.  
The formula of closeness value is:  $\text{Max}(\text{Value1}, \text{Value2}) / (\text{Value1} + \text{Value2})$ .

\*\*\*\*\* Closeness Matrix of 6 tagged files \*\*\*\*\*

Tagged Files	F 2	F 3	F 4	F 5	F 6
AP880328-0206.tagged [F 1]	11/14	5/7	1/2	10/13	3/5
AP891110-0043.tagged [F 2]		22/37	11/14	11/21	22/31
AP890925-0054.tagged [F 3]			5/7	4/7	5/8
AP900220-0065.tagged [F 4]				10/13	3/5
FT923-5797.tagged [F 5]					20/29
FT923-5835.tagged [F 6]					



From the above five individual analysis reports, we can derive the following reports:

(Each file is compared with the rest of files with different weight percentage value)

**Report-1 (File-1 compared with the rest of files)**

F1	F2	F3	F4	F5	F6
0 %	79.4	73.1	53.3	78.8	58.8
20 %	77.4	70.8	53.3	75.0	56.3
40 %	78.6	71.4	50.0	76.9	60.0
Max (Avg)	78.5				

**Report-2 (File-2 compared with the rest of files)**

F2	F3	F4	F5	F6
0 %	58.7	77.1	50.9	73.0
20 %	58.5	75.0	55.8	72.7
40 %	59.4	78.6	52.4	71.0
Max (Avg)		76.9		

**Report-3 (File-3 compared with the rest of files)**

F3	F4	F5	F6
0 %	70.4	57.8	65.5
20 %	68.0	55.3	65.4
40 %	71.4	57.1	62.5
Max (Avg)	69.9		

#### Report-4 (File-4 compared with the rest of files)

F4	F5	F6
0 %	76.5	55.6
20 %	72.4	52.9
40 %	76.9	60.0
Max (Avg)	75.3	

#### Report-5 (File-5 compare with the rest files)

F5	F6
0 %	72.2
20 %	70.0
40 %	69.0
Max (Avg)	70.4

For all the above the five derivate reports, we get the largest average closeness value for each report, and based on this we can have the chain as: F1-F2 (78.5), F2-F4 (76.9), F5-F6 (70.4) and F3-F4 (69.9). We use the same way as optimizing the draft clusters and we can get: F1-F2-F4 are the closest three documents, F5-F6 are the closest two documents. Fortunately, document-1, document-2 and document-4 all are about entertainment, and document-5 and document-6 are about economy. So far the results are promising.

We also have F3-F4 (69.9). Document-3 is about natural disaster. It should not be closed to document-4 which is about entertainment. But if we trace back to single document analysis part (see Page 36), there we have already analyzed that *storm*, *hurricane*, *damage*, *injury*, etc. should be in a cluster, and they all should be in the largest cluster and then we will have a

lower closeness value because F3-F4 actually are not in different groups only because multiple senses without non-disambiguation.

So, due to the multiple senses of non-disambiguation problem in WordNet, it is possible to cluster those non-related nouns. This is also referred in [Inkpen, Diana Zaiu and Hirst Graeme, 2001] *“There are too many senses for each word (for example, six senses for error, eight for absorb). We need to disambiguate to see what are the senses closely related to the peripheral concept. We have to group together senses which are very similar. The WordNet hierarchy has to be reorganized to accommodate the clusters of near-synonyms. From this point of view, Mikrokosmos [Kavi Mahesh and Sergei Nirenburg, 1995] is better because there are few senses for a word; closely related senses were merged into one sense when the ontology was built.”*

## **5. Conclusion and future work**

Through this project, we developed a simple distance measure on WordNet and showed how it can be used to determine the semantically related nouns to be clustered. Further we can compare closeness between every two documents based on their clusters of nouns. And the evaluation against human judgements shows the effectiveness of the resulting measures.

Even though the result showed in this report is not perfect, we can conclude that it is a very interesting attempt and a possible approach.

In this project we simply cluster the nouns without considering word sense disambiguation.

If we have an intermediate model that can make a decision to choose the proper sense of a word through context analysis, it will reduce the invalid clusters and all those invalid clustered nouns will be removed.

## References

1. Sabine Bergler 1993. Semantic Dimensions in the Field of Reporting Verbs, in *Proceedings of the Ninth Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary and Text Research*, Oxford, England, September 1993.
2. Budanitsky, Alexander and Hirst Graeme 2001. `Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, June 2001.
3. Inkpen, Diana Zaiu and Hirst Graeme 2001. Building a lexical knowledge-base of near-synonym differences. *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, June 2001.
4. The EAGLES Lexicon Interest Group. 1999. Preliminary recommendations on lexical semantics encoding. Final report EAGLES LE3-4244, EAGLES, February.
5. Brian Harvey and Matthew Wright 1994. *Simply Scheme*. The MIT Press.
6. D. Lin, 1998. An Information-Theoretic Definition of Similarity. *Proceedings of International Conference on Machine Learning*, Madison, Wisconsin, July, 1998.

7. Miller, G. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography* 3 (4). The MIT Press.
8. Gary Olsen, University of Illinois. Some lecture notes on Classification and Phylogeny 2003. (<http://geta.life.uiuc.edu/~gary/>)
9. Kavi Mahesh and Sergei Nirenburg. 1995. A situated ontology for practical NLP. In *Proceedings of Workshop on Basic Ontological Issues in Knowledge Sharing, International JointConference on Artificial Intelligence*, Montreal, Canada.

## Appendix A. Scheme Code Samples

### Example-1

The procedure to generate the matrix of semantic relations between each two words is as following:

```
(define create_matrix
  (lambda ()
    (set! wd1 (car (list-ref nn_cate_list curr_nn_idx)))
    (set! wd2 (car (list-ref nn_cate_list next_nn_idx)))

    (set! max_cate_of_curr_nn (- (length (cdr (list-ref nn_cate_list curr_nn_idx))) 1))
    (set! max_cate_of_next_nn (- (length (cdr (list-ref nn_cate_list next_nn_idx))) 1))
    (set! curr_nn_curr_cate_idx 0)
    (set! next_nn_curr_cate_idx 0)

    (set! count_of_cate_matched_4_matrix 0)
    (create_matrix_4_these_2_wds)

    (if (eqv? bool_dispdot #t) (display "."))
    (set! idx (+ idx 1))

    ;here,check highest_matrix_value only if they are not the same meanings ;e.g. Job/job,
    Job/jobs, job/jobs, Jobs/job
    (if (or (eqv? (string-ci=? (car wd1) (car wd2)) #t)
            (eqv? (string-ci=? (string-append (car wd1) "s") (car wd2)) #t) (eqv? (string-ci=? (car
wd1) (string-append (car wd2) "s")) #t))
        (display "")
        (begin
          (if (< highest_matrix_value count_of_cate_matched_4_matrix)
              (set! highest_matrix_value count_of_cate_matched_4_matrix))))

    (output_matrix_content wd1 wd2 count_of_cate_matched_4_matrix)

    (if (equal? matrix_list '())
        (set! matrix_list (list (cons wd1 (list (cons wd2
count_of_cate_matched_4_matrix))))))
        (begin
          (set! matrix_list (append matrix_list (list (cons wd1 (list (cons
wd2
count_of_cate_matched_4_matrix))))))))))
```

```

(if (<= (- max_nn_wds next_nn_idx) 1)
  (if (<= (- max_nn_wds curr_nn_idx) 2)
    (begin
      (newline result_port)
      (display "      " result_port)
      (list_dash_ln_in_header 1 (* (- max_nn_wds 1) 5))
      (newline result_port)
      (display "Notes:" result_port)
      (print_notes 0)
      (newline)
      (display "Clustering noun words from matrix      ....")
      (set! matrix_list_length (length matrix_list))
      (set! loopidx 0)
      (set! matrix_prevwd (caar (list-ref matrix_list 0)))
      (set! temp_clustered_list '())

      ;//before process organize_matrix,
      ;//user entered percentage value will be re-calculated here
      (set! cluster_weight (/ (* cluster_weight highest_matrix_value) 100))
      (organize_matrix matrix_prevwd temp_clustered_list loopidx)
      (newline)
      (display write_msg)
      (newline)
      (newline)

      ;temp for printing clustered wds into file
      (newline result_port)
      (newline result_port)
      (display "-----" result_port)
      (newline result_port)
      (display (string-append "Clustered words (weight percentage : "
        (number->string cluster_percentage)
        " % ) :") result_port)
      (display_result 0 (length clustered_wds_list))
      (close-output-port result_port))

    (begin
      (set! curr_nn_idx (+ curr_nn_idx 1))
      (set! next_nn_idx (+ curr_nn_idx 1))
      (create_matrix)))
  (begin
    (set! next_nn_idx (+ next_nn_idx 1))
    (create_matrix))))

```



## Example-2

The procedures to generate the matrix of title and matrix structure are as following:

```
(define list_all_wd_in_header
  (lambda (idx length_nn_wds)
    (if (< idx length_nn_wds)
        (begin
          (if (< idx 3)
              (display (string-append "      Wd" (number->string idx) " ") result_port)
              (cond
                ((< idx 10)
                 (display (string-append "Wd" (number->string idx) " ") result_port))
                ((< idx 100)
                 (display (string-append "W" (number->string idx) " ") result_port))
                ((< idx 1000)
                 (display (string-append (number->string idx) " ") result_port))))
          (set! idx (+ idx 1))
          (list_all_wd_in_header idx length_nn_wds))))))
```

; print output header

```
(define list_dash_ln_in_header
  (lambda (idx max_dash)
    (if (<= idx max_dash)
        (begin
          (display "-" result_port)
          (set! idx (+ idx 1))
          (list_dash_ln_in_header idx max_dash))))))
```

;//print detailed matrix //

```
(define output_matrix_content
  (lambda (wd1 wd2 matched_count)
    (if (not (equal? wd1 matrix_compare_wd))
        (begin
          (set! matrix_compare_wd wd1)
          (newline result_port)
          (cond
            ((< idx 10)
             (display (string-append "Wd" (number->string idx) " -->") result_port))
            ((< idx 100)
             (display (string-append "W" (number->string idx) " -->") result_port))
            ((< idx 1000)
             (display (string-append (number->string idx) " -->") result_port)))
          (print_space 1 (* (- idx 1) 5)))
```

```

        (display (string-append " " (number->string matched_count) " ") result_port))
    (begin
      (set! idx (- idx 1))
      (if (< matched_count 10)
        (display (string-append (number->string matched_count) " ") result_port)
        (display (string-append (number->string matched_count) " ") result_port))))
  )
)

```

```

(define print_space
  (lambda (counter maxspace)
    (if (<= counter maxspace)
      (begin
        (display " " result_port)
        (print_space (+ counter 1) maxspace))))))

```

## Appendix B. Contents of Tagged Files

### Content of tagged file AP880328-0206.tagged

<s>/SYM <DOC>/SYM  
<DOCNO>/SYM AP880328-0206/NNP </DOCNO>/SYM  
<FILEID>AP-NR-03-28-88/JJ 1345EST</FILEID>/CD  
<FIRST>s/NNS e/NNP BC-APN--Oscars-Questions/NNP Adv10/NNP 03-28/JJ  
0411</FIRST>/CD  
<SECOND>BC-APN--Oscars-Questions/NNS ,/, Adv/NNP 10,0432</SECOND>/CD  
<NOTE>\$adv10</NOTE>/JJ  
<HEAD>AGENCIES/NNS AND/CC RADIO/NNP OUT</HEAD>/NNP  
<NOTE>For/NNP Release/NNP Sunday/NNP ,/, April/NNP 10</NOTE>/CD  
<HEAD>From/NNP AP/NNP Newsfeatures</HEAD>/NNP  
<HEAD>With/NNP BC-APN--Oscars</HEAD>/NNP  
<TEXT>/NN  
HOLLYWOOD/NNP (AP)/NN \_/NN Like/IN all/DT great/JJ contests/NNS ,/, the/DT  
annual/JJ Academy/NNP  
Awards/NNP ceremony/NN has/VBZ generated/VBN a/DT wealth/NN of/IN  
statistics/NNS and/CC trivia/NNS  
over/IN its/PRP\$ 60/CD years/NNS of/IN existence/NN ,/, all/DT carefully/RB  
chronicled/VBN by/IN Oscar/NNP  
historians./JJ  
Here/RB ,/, courtesy/NN of/IN the/DT Academy/NNP of/IN Motion/NNP  
Picture/NNP Arts/NNP and/CC  
Sciences/NNPS ,/, are/VBP 20/CD questions/NNS about/IN Oscar:/NNP  
1./CD What/WP actress/NN has/VBZ received/VBN the/DT most/RBS Oscars?/NNP  
Katharine/NNP Hepburn,/NNP  
4./CD  
2./CD What/WP actor/NN has/VBZ received/VBN the/DT most/RBS Oscars?/NNP  
Walter/NNP Brennan/NNP ,/, 3./CD  
3./CD What/WP director/NN has/VBZ received/VBN the/DT most/RBS Oscars?/NNP  
John/NNP Ford/NNP ,/, 4./CD  
4./CD Who/WP is/VBZ the/DT most/RBS nominated/VBN actress?/NNP Katharine/NNP  
Hepburn/NNP ,/, 12./CD  
5./CD Who/WP is/VBZ the/DT most/RBS nominated/VBN actor?/NNP Laurence/NNP  
Olivier/NNP ,/, 10./CD  
6./CD Who/WP is/VBZ the/DT most/RBS nominated/VBN director?/NNP William/NNP  
Wyler/NNP ,/, 12./CD  
7./CD How/WRB long/JJ was/VBD the/DT longest/JJS Oscar/NNP ceremony?/NN 3/CD  
hours/NNS ,/, 45/CD minutes/NNS  
in/IN 1985./CD  
8./CD What/WP film/NN has/VBZ won/VBN the/DT most/RBS Oscars?/NNP ```` Ben-Hur,``/NNP 11/CD ,/, 1959./CD  
9./CD What/WP was/VBD the/DT most/RBS nominated/VBN picture?/NN ```` Gone/NNP With/IN The/DT Wind,``/NNP  
```` Mary/NNP Poppins``/NNP and/CC ```` Who/WP 's/VBZ Afraid/JJ  
of/IN Virginia/NNP Woolf,``/NNP each/DT with/IN  
13./CD  
10./CD What/WP studio/NN has/VBZ won/VBN the/DT most/RBS Oscars?/NNP MGM/NNP  
,/, with/IN nearly/RB 200./CD  
11./CD Who/WP was/VBD the/DT youngest/JJS performer/NN to/TO win/VB an/DT  
Oscar?/NNP Tatum/NNP  
O'Neal/NNP ,/, 10/CD ,/, supporting/VBG actress/NN ,/, ```` Paper/NNP  
Moon,``/NNP 1973./CD

12./CD Who/WP was/VBD the/DT oldest/JJS performer/NN ever/RB to/TO win/VB  
 an/DT Oscar?/NNP Groucho/NNP  
 Marx/NNP ,/, 83/CD ,/, an/DT honorary/JJ award/NN in/IN 1974./CD George/NNP  
 Burns/NNP is/VBZ the/DT oldest/JJS  
 actor/NN to/TO win/VB a/DT performing/VBG Oscar/NNP ,/, 80/CD ,/, for/IN  
 the/DT `````` The/DT Sunshine/NNP Boys''/NNP  
 in/IN 1976./CD  
 13./CD What/WP actor/NN won/VBD a/DT posthumous/JJ award?/NNP Peter/NNP  
 Finch/NNP ,/, `````` Network, ''/NNP  
 1976./CD  
 14./CD What/WP father-son/NN team/NN won/VBD Oscars/NNP in/IN the/DT same/JJ  
 picture?/NNP Walter/NNP  
 Huston/NNP ,/, supporting/VBG actor/NN ,/, and/CC John/NNP Huston/NNP ,/,  
 director-writer/JJ ,/, `````` The/DT  
 Treasure/NNP of/IN Sierra/NNP Madre, ''/NNP 1948./CD  
 15./CD What/WP actor/NN was/VBD nominated/VBN in/IN two/CD categories/NNS  
 for/IN the/DT same/JJ  
 role?/NNP Barry/NNP Fitzgerald/NNP ,/, `````` Going/VBG My/PRP\$  
 Way, ''/NNP 1944./CD He/PRP won/VBD in/IN the/DT  
 supporting/VBG category./JJ  
 16./CD Who/WP decapitated/VBD his/PRP\$ Oscar/NNP while/IN practicing/VBG  
 golf/NN swings/NNS  
 indoors?/NNP Barry/NNP Fitzgerald./NNP His/PRP\$ 1944/CD supporting/VBG  
 actor/NN statuette/NN was/VBD  
 subsequently/RB replaced/VBN by/IN the/DT Academy./NNP  
 17./CD What/WP performers/NNS have/VBP refused/VBN their/PRP\$ Oscars?/NNP  
 George/NNP C./NNP Scott/NNP  
 and/CC Marlon/NNP Brando./NNP  
 18./CD How/WRB many/JJ people/NNS view/VBP the/DT Oscar/NNP ceremony/NN  
 on/IN television?/NN Last/JJ  
 year/NN ,/, the/DT Oscarcast/NNP was/VBD seen/VBN by/IN one/CD billion/CD  
 people/NNS in/IN 79/CD countries./JJ  
 19./CD Who/WP holds/VBZ the/DT record/NN for/IN the/DT fastest/RBS  
 televised/VBN reading/NN of/IN  
 the/DT famed/JJ Academy/NNP rules?/NNP Actor-comedian/NNP John/NNP  
 (Federal/NNP Express)/NNP  
 Moschitta/NNP ,/, 25/CD seconds/NNS ,/, 1983./CD  
 20./CD What/WP multiple/JJ Oscar/NNP winner/NN has/VBZ never/RB appeared/VBN  
 to/TO collect/VB her/PRP\$  
 award?/NNP Katharine/NNP Hepburn/NNP ,/, all/DT four/CD times./JJ  
 </TEXT>/NN  
 <NOTE>END/NNP ADV/NNP for/IN Sunday/NNP ,/, April/NNP 10</NOTE>/CD  
 </DOC>/SYM

## Content of tagged file AP891110-0043.tagged

<s>/SYM <DOC>/SYM  
<DOCNO>/SYM AP891110-0043/CD </DOCNO>/SYM  
<FILEID>AP-NR-11-10-89/CD 0356EST</FILEID>/CD  
<FIRST>r/VBG a/DT PM-OscarCopyright/NNP 11-10/JJ 0359</FIRST>/CD  
<SECOND>PM-Oscar/NNP Copyright,0368</SECOND>/NNP  
<HEAD>Oscar/NNP Trophy/NNP Has/VBZ No/DT Copyright</HEAD>/NNP  
<BYLINE>By/NNP JOHN/NNP HORN</BYLINE>/NNP  
<BYLINE>Associated/VBN Press/NNP Writer</BYLINE>/NNP  
<DATELINE>LOS/NNP ANGELES/NNP (AP)/NN </DATELINE>/SYM  
<TEXT>/NN  
The/DT Oscar/NNP statuette/NN ,/, one/CD of/IN the/DT most/RBS  
recognizable/JJ images/NNS in/IN the/DT entertainment/NN world/NN ,/,  
has/VBZ no/DT copyright/NN  
protection/NN ,/, a/DT federal/JJ judge/NN has/VBZ ruled./JJ  
The/DT small/JJ Academy/NNP Award/NNP statue/NN is/VBZ part/NN of/IN the/DT  
public/JJ domain,/NN  
U.S./NNP District/NNP Court/NNP Judge/NNP Laughlin/NNP Waters/NNP said/VBD  
in/IN a/DT ruling/NN released/VBN  
Thursday./NNP  
The/DT decision/NN was/VBD a/DT setback/NN for/IN the/DT Academy/NNP of/IN  
Motion/NNP Picture/NNP  
Arts/NNP and/CC Sciences/NNPS ,/, which/WDT had/VBD sued/VBN a/DT Chicago-  
based/JJ manufacturer/NN of/IN  
an/DT employee-incentive/NN trophy/NN similar/JJ to/TO the/DT Oscar./NNP  
Academy/NNP President/NNP Karl/NNP Malden/NNP said/VBD the/DT ruling/NN ``  
`` comes/VBZ as/IN a/DT shock/NN  
to/TO me.''/CD The/DT academy/NN said/VBD it/PRP would/MD appeal./CD  
The/DT academy/NN claimed/VBD that/IN the/DT Star/NNP Award/NNP ,/, the/DT  
trophy/NN look-alike/JJ  
made/VBN by/IN Creative/NNP House/NNP Promotions/NNS ,/, violated/VBD  
copyright/NN laws/NNS ,/, diluted/VBN  
the/DT academy/NN 's/POS trademark/NN and/CC represented/VBN unfair/JJ  
competition./JJ  
The/DT Star/NNP Award/NNP depicted/VBD a/DT naked/JJ ,/, muscular/JJ male/JJ  
much/NN like/IN the/DT  
Oscar/NNP ,/, just/RB two/CD inches/NNS shorter/JJR and/CC holding/VBG a/DT  
star/NN instead/RB of/IN a/DT  
sword./JJ It/PRP had/VBD a/DT gold/NN finish/NN similar/JJ to/TO the/DT  
Oscar/NNP ,/, and/CC stood/VBD on/IN a/DT  
circular/JJ gold/NN cap/NN mounted/VBN on/IN a/DT cylindrical/JJ base./CD  
Although/IN Waters/NNP acknowledged/VBD that/IN the/DT Creative/NNP  
House/NNP trophy/NN is/VBZ  
`` `` very/JJ similar''/NN to/TO the/DT Oscar/NNP ,/, he/PRP  
rejected/VBD all/DT of/IN the/DT academy/NN 's/POS  
legal/JJ claims/NNS ,/, saying/VBG that/IN the/DT statuette/NN became/VBD  
part/NN of/IN the/DT public/JJ  
domain/NN prior/RB to/TO Jan./NNP 1/CD ,/, 1978/CD ,/, the/DT effective/JJ  
date/NN of/IN the/DT Copyright/NNP  
Act/NNP of/IN 1976./CD  
The/DT film/NN industry/NN 's/POS top/JJ award/NN was/VBD distributed/VBN  
from/IN the/DT first/JJ  
ceremony/NN in/IN 1929/CD until/IN about/IN 1941/CD without/IN the/DT ``  
`` c''/NN mark/NN indicating/VBG  
a/DT copyright./JJ

The/DT judge/NN ruled/VBD ,/, too/RB ,/, that/IN the/DT Oscar/NNP not/RB  
only/RB was/VBD an/DT honorary/JJ  
award/NN but/CC also/RB was/VBD used/VBN to/TO promote/VB the/DT film/NN  
industry/NN ,/, and/CC therefore/RB  
did/VBD not/RB have/VB a/DT `` `` `` limited/JJ purpose,``/NN as/IN  
required/VBN by/IN the/DT copyright/NN  
act./JJ  
The/DT academy/NN noted/VBD that/IN the/DT ruling/NN applied/VBN only/RB  
to/TO the/DT statuette/NN  
and/CC that/IN the/DT academy/NN 's/POS trademark/NN and/CC service/NN  
marks/NNS for/IN the/DT Oscar/NNP  
and/CC the/DT Academy/NNP Award/NNP names/NNS and/CC symbols/NNS are/VBP  
unaffected./JJ  
`` `` `` We/PRP are/VBP surprised/VBN that/IN the/DT court/NN would/MD  
base/VB its/PRP\$ decision/NN on/IN  
events/NNS which/WDT occurred/VBD half/PDT a/DT century/NN ago/RB ,/,  
overlooking/VBG the/DT  
meticulous/JJ fashion/NN in/IN which/WDT the/DT Academy/NNP has/VBZ  
protected/VBN the/DT integrity/NN  
of/IN its/PRP\$ institutional/JJ symbol/NN for/IN decades,``/NNP Malden/NNP  
said/VBD in/IN a/DT  
statement./JJ  
</TEXT>/NN  
</DOC>/SYM

## Content of tagged file AP890925-0054.tagged

<s>/SYM <DOC>/SYM  
<DOCNO>/SYM AP890925-0054/CD </DOCNO>/SYM  
<FILEID>AP-NR-09-25-89/CD 0631EDT</FILEID>/CD  
<FIRST>r/VBG a/DT PM-Hugo-Readiness/NNP 09-25/CD 0321</FIRST>/CD  
<SECOND>PM-Hugo-Readiness,0336</SECOND>/JJ  
<HEAD>Loss/NN Of/IN Life/NNP Low/NNP Because/IN People/NNS Were/VBD  
Prepared</HEAD>/NNP  
<HEAD>With/NNP PM-Hugo/NNP Bjt</HEAD>/NNP  
<BYLINE>By/NNP BRUCE/NNP SMITH</BYLINE>/NNP  
<BYLINE>Associated/VBN Press/NNP Writer</BYLINE>/NNP  
<DATELINE>CHARLESTON/NN ,/, S.C./NNP (AP)/NN </DATELINE>/SYM  
<TEXT>/NN  
Residents/NNS took/VBD forecasters/NNS at/IN their/PRP\$ word/NN  
when/WRB they/PRP warned/VBD of/IN Hurricane/NNP Hugo/NNP 's/POS fury/NN ,/,  
and/CC the/DT low/JJ number/NN of/IN  
deaths/NNS from/IN the/DT powerful/JJ storm/NN can/MD be/VB credited/VBN  
to/TO this/DT healthy/JJ  
respect/NN ,/, authorities/NNS said./JJ  
The/DT storm/NN ,/, which/WDT caused/VBD billions/NNS in/IN damage/NN ,/,  
claimed/VBD 17/CD lives/NNS in/IN  
South/NNP Carolina/NNP ,/, and/CC only/RB two/CD were/VBD in/IN the/DT  
Charleston/NNP area/NN ,/, which/WDT  
bore/VBD the/DT brunt/NN of/IN Hugo/NNP 's/POS 135/CD mph/NN winds./JJ  
'/'/'/'/'/' We/PRP just/RB feel/VBP very/RB thankful/JJ about/IN that,''/NN  
said/VBD Mayor/NNP Joseph/NNP P./NNP  
Riley/NNP Jr./NNP '/'/'/'/'/' Several/JJ thousands/NNS of/IN people/NNS who/WP  
were/VBD in/IN the/DT shelters/NNS  
and/CC the/DT tens/NNS of/IN thousands/NNS of/IN people/NNS who/WP  
evacuated/VBN inland/RB were/VBD  
potential/JJ victims/NNS of/IN injury/NN and/CC death.''/JJ  
Riley/NNP and/CC other/JJ officials/NNS credited/VBD residents/NNS with/IN  
monitoring/VBG the/DT  
hurricane/NN as/IN it/PRP spun/VBD toward/IN Charleston/NNP ,/, and/CC  
heeding/VBG evacuation/NN  
warnings./CD  
The/DT last/JJ big/JJ storm/NN to/TO hit/VB South/NNP Carolina/NNP was/VBD  
Hurricane/NNP David/NNP in/IN  
1979/CD ,/, which/WDT hit/VBD the/DT coast/NN after/IN pounding/VBG  
southern/JJ Florida./NNP  
Since/IN then/RB several/JJ big/JJ storms/NNS have/VBP threatened/VBN but/CC  
veered/VBD away,/NN  
and/CC officials/NNS were/VBD worried/VBN that/IN Charlestonians/NNPS  
would/MD n't/RB heed/VB Hugo/NNP 's/POS  
warnings./CD  
'/'/'/'/'/' We/PRP 've/VBP had/VBD so/RB many/JJ close/NN calls,''/NN said/VBD  
Gary/NNP Garnet/NNP ,/, a/DT  
meteorologist/NN with/IN the/DT National/NNP Weather/NNP Service./NNP  
'/'/'/'/'/' But/CC this/DT storm/NN was/VBD very/RB ,/, very/JJ strong,''/NN  
he/PRP said./JJ '/'/'/'/'/' This/DT time/NN  
they/PRP realized/VBD it.''/JJ  
Glen/NNP Ellis/NNP ,/, a/DT Red/NNP Cross/NNP official/NN ,/, estimated/VBN  
that/DT 18,000/CD  
residents/NNS in/IN Charleston/NNP and/CC the/DT three/CD surrounding/VBG  
counties/NNS headed/VBN

for/IN shelters./JJ  
Tens/NNS of/IN thousands/NNS more/JJR got/VBD into/IN their/PRP\$ cars/NNS  
and/CC headed/VBN up/IN  
Interstate/NNP 26/CD toward/IN Columbia./NNP  
Mayor/NNP John/NNP Bourne/NNP of/IN North/NNP Charleston/NNP ,/, where/WRB  
there/EX apparently/RB  
were/VBD no/DT storm-related/JJ deaths/NNS ,/, credited/VBN people/NNS  
for/IN heeding/VBG the/DT  
warnings./CD  
`/`` `/'` It/PRP was/VBD bad/JJ ,/, but/CC it/PRP could/MD have/VB been/VBN  
a/DT lot/NN worse,``/NN he/PRP said./JJ  
`/`` `/'` You/PRP 've/VBP just/RB got/VBD to/TO say/VB nice/JJ things/NNS  
about/IN the/DT people.``/CD  
Meanwhile/RB ,/, Garnet/NNP had/VBD some/DT good/JJ news/NN for/IN storm-  
weary/JJ  
Charlestonians/NNS :: late/JJ Saturday/NNP the/DT tropics/NNS were/VBD  
quiet/JJ with/IN no/DT other/JJ  
storms/NNS brewing./CD  
`/`` `/'` There/EX 's/VBZ nothing/NN to/TO worry/VB about/IN out/IN  
there,``/NN he/PRP said./JJ  
</TEXT>/NN  
</DOC>/SYM



## Content of tagged file AP900220-0065.tagged

<s>/SYM <DOC>/SYM  
<DOCNO>/SYM AP900220-0065/CD </DOCNO>/SYM  
<FILEID>AP-NR-02-20-90/CD 1033EST</FILEID>/CD  
<FIRST>r/VBG a/DT PM-People-SpikeLee/NNP 02-20/JJ 0218</FIRST>/CD  
<SECOND>PM-People-Spike/NNP Lee,0223</SECOND>/NNP  
<HEAD>Director/NNP Spike/NNP Lee/NNP Has/VBZ Pointed/JJ Remarks/NNS in/IN  
Talk/NN at/IN Syracuse/NNP  
University</HEAD>/NNP  
<DATELINE>SYRACUSE/NN ,/, N.Y./NNP (AP)/NN </DATELINE>/SYM  
<TEXT>/NN  
Spike/NNP Lee/NNP may/MD have/VB been/VBN wearing/VBG a/DT  
Georgetown/NNP University/NNP cap/NN and/CC sweatshirt/NN ,/, but/CC he/PRP  
got/VBD a/DT far/RB warmer/JJR  
reception/NN from/IN Syracuse/NNP University/NNP students/NNS than/IN he/PRP  
did/VBD from/IN  
Academy/NNP Award/NNP voters./JJ  
In/IN a/DT lecture/NN and/CC question-and-answer/JJ session/NN Monday/NNP  
,/, Lee/NNP  
shrugged/VBD off/IN the/DT academy/NN 's/POS icy/JJ treatment/NN of/IN  
his/PRP\$ latest/JJS movie/NN ,/, `` `` `` Do/VBP  
the/DT Right/RB Thing,''/NNP and/CC blamed/VBD it/PRP on/IN generational/JJ  
politics./JJ  
The/DT 1989/CD film/NN ,/, widely/RB hailed/VBN by/IN critics/NNS ,/,  
received/VBD only/RB two/CD  
Academy/NNP Award/NNP nominations/NNS ,/, and/CC is/VBZ n't/RB being/VBG  
considered/VBN for/IN best/JJS  
picture/NN or/CC best/JJS director./JJ  
The/DT academy/NN 's/POS membership/NN consists/VBZ mainly/RB of/IN `` ``  
`` `` old/JJ people,''/VBG the/DT  
32-year-old/NNP Lee/NNP said./JJ He/PRP said/VBD they/PRP were/VBD `` `` ``  
a/DT lot/NN more/RBR comfortable/JJ  
with/IN the/DT black/JJ chauffeur/NN in/IN `` `` Driving/VBG Miss/NNP  
Daisy'''/NNP than/IN with/IN the/DT  
angry/JJ characters/NNS in/IN his/PRP\$ film./CD  
`` `` `` I/PRP still/RB feel/VBP we/PRP made/VBD the/DT best/JJS film/NN  
of/IN the/DT year,''/NN he/PRP said./JJ  
Asked/VBN if/IN racism/NN might/MD have/VB been/VBN involved/VBN in/IN  
the/DT academy/NN  
nominations/NNS ,/, Lee/NNP replied/VBD ,/, `` `` `` Racism/NN 's/POS  
involved/VBN in/IN everything.''/JJ  
Despite/IN the/DT cap/NN and/CC sweatshirt/NN ,/, Syracuse/NNP students/NNS  
welcomed/VBD him/PRP  
as/IN one/CD of/IN their/PRP\$ own./CD  
`` `` `` He/PRP 's/VBZ an/DT African-American/NNP ,/, one/CD of/IN the/DT  
first/JJ ones/NNS to/TO make/VB it/PRP (as/VBZ  
a/DT producer/NN and/CC director)/NN ,/, and/CC he/PRP 's/VBZ addressing/VBG  
the/DT issues/NNS and/CC  
concerns/NNS affecting/VBG our/PRP\$ community,''/NNP Syracuse/NNP senior/JJ  
Deron/NNP Harris/NNP  
said./JJ `` `` `` That/DT 's/VBZ why/WRB we/PRP 're/VBP so/RB proud/JJ  
of/IN him.''/CD  
</TEXT>/NN  
</DOC>/SYM

## Content of tagged file FT923-5797.tagged

<s>/SYM <DOC>/SYM  
<DOCNO>FT923-5797</DOCNO>/CD  
<PROFILE>\_AN-CIBBCACSFT</PROFILE>/JJ  
<DATE>920828/CD  
</DATE>/NN  
<HEADLINE>/NN  
FT/NNP 28/CD AUG/NNP 92/CD //NN Cleaning/VBG up/IN after/IN Andrew/NNP  
</HEADLINE>/NN  
<BYLINE>/NN  
By/IN AGENCIES/NNS  
</BYLINE>/NN  
<DATELINE>/SYM  
FLORIDA/NNP ,/, NEW/NNP ORLEANS/NNP  
</DATELINE>/SYM  
<TEXT>/NN  
SQUADS/NNS of/IN workers/NNS fanned/VBD out/IN across/IN storm-battered/NNP  
Louisiana/NNP yesterday/NN to/TO  
begin/VB a/DT massive/JJ rebuilding/NN effort/NN after/IN Hurricane/NNP  
Andrew/NNP had/VBD flattened/VBN whole/JJ  
districts/NNS ,/, killing/VBG two/CD people/NNS and/CC injuring/VBG  
dozens/NNS more/JJR ,/, agencies/NNS report/VBP from/IN  
Florida/NNP and/CC New/NNP Orleans./NNP  
However/RB ,/, local/JJ officials/NNS in/IN Florida/NNP ,/, hit/VBD  
earlier/RBR in/IN the/DT week/NN by/IN the/DT  
hurricane/NN ,/, were/VBD critical/JJ of/IN what/WP they/PRP called/VBD a/DT  
delay/NN in/IN supplying/VBG food,/NN  
drinking/NN water/NN and/CC other/JJ supplies/NNS for/IN thousands/NNS of/IN  
people/NNS in/IN need./JJ  
Federal/JJ emergency/NN officials/NNS acknowledged/VBD distribution/NN  
problems,/NN  
Transportation/NNP Secretary/NNP Andrew/NNP Card/NNP yesterday/NN  
promised/VBD 'dramatic'/NN  
improvements/NNS within/IN 24/CD hours/NNS and/CC President/NNP George/NNP  
Bush/NNP last/JJ night/NN ordered/VBD  
troops/NNS to/TO Florida/NNP ,/, without/IN specifying/VBG a/DT number./CD  
The/DT government/NN estimated/VBD it/PRP would/MD cost/VB Dollars/NNPS  
20bn-Dollars/CD 30bn/CD to/TO tidy/JJ and/CC  
rebuild/VB in/IN Florida/NNP ,/, and/CC to/TO care/VB for/IN residents/NNS  
displaced/VBN by/IN the/DT storm./JJ  
Louisiana/NNP state/NN officials/NNS said/VBD they/PRP had/VBD no/DT  
overall/JJ count/NN of/IN storm-related/JJ  
injuries/NNS but/CC initial/JJ estimates/NNS reckoned/VBN fewer/JJR than/IN  
100./CD The/DT Federal/NNP  
Emergency/NNP Management/NNP Agency/NNP said/VBD it/PRP was/VBD setting/VBG  
aside/RB Dollars/NNPS 77m/CD to/TO help/VB  
Louisiana/NNP recover./CD  
Most/JJS of/IN the/DT storm/NN 's/POS fury/NN was/VBD spent/VBN against/IN  
sparsely/RB populated/VBN farming/NN  
communities/NNS and/CC swampland/NN in/IN the/DT state/NN ,/, sparing/VBG  
it/PRP the/DT widespread/JJ  
destruction/NN caused/VBN in/IN Florida/NNP ,/, where/WRB 15/CD people/NNS  
died./JJ  
Official/JJ estimates/NNS in/IN Miami/NNP reported/VBD that/IN the/DT  
hurricane/NN had/VBD wiped/VBN out/RP the/DT

homes/NNS of/IN one/CD Dade/NNP County/NNP resident/NN in/IN eight/CD -/:  
a/DT quarter/NN of/IN a/DT million/CD people./CD  
Andrew/NNP had/VBD become/VBN little/RB more/JJR than/IN a/DT strong/JJ  
rainstorm/NN early/JJ yesterday,/NN  
moving/VBG across/IN Mississippi/NNP state/NN and/CC heading/VBG for/IN  
the/DT north-eastern/NNP US./NNP  
Several/JJ of/IN Louisiana/NNP 's/POS main/JJ industries/NNS were/VBD  
affected/VBN ,/, including/VBG those/DT of/IN  
oysters/NNS and/CC alligators./NNP Wildlife/NNP and/CC fisheries/NNS  
secretary/NN Joe/NNP Herring/NNP  
estimated/VBN a/DT 50/CD per/IN cent/NN decline/NN in/IN the/DT alligator/NN  
industry./VBD The/DT cotton/NN and/CC  
sugar-cane/JJ crops/NNS were/VBD threatened/VBN ,/, the/DT state/NN  
agriculture/NN department/NN said./JJ  
Most/JJS Louisiana/NNP oil/NN refineries/NNS ,/, however/RB ,/, were/VBD  
barely/RB affected/VBN and/CC deliveries/NNS  
of/IN crude/JJ oil/NN were/VBD expected/VBN to/TO resume/VB yesterday./JJ  
</TEXT>/NN  
<PUB>The/NNP Financial/NNP Times/NNP  
</PUB>/NN  
<PAGE>/NN  
London/NNP Page/NNP 4/CD  
</PAGE>/NN  
</DOC>/SYM

## Content of tagged file FT923-5835.tagged

<s>/SYM <DOC>/SYM  
<DOCNO>FT923-5835</DOCNO>/CD  
<PROFILE>\_AN-CIBBCABPFT</PROFILE>/JJ  
<DATE>920828/CD  
</DATE>/NN  
<HEADLINE>/NN  
FT/NNP 28/CD AUG/NNP 92/CD //NN UK/NNP Company/NNP News/NNP :/: GA/NNP  
says/VBZ hurricane/NN claims/NNS could/MD reach/VB 'up/NN to/TO  
Dollars/NNS 40m'/CD  
</HEADLINE>/NN  
<BYLINE>/NN  
By/IN ROBERT/NNP PESTON/NNP  
</BYLINE>/NN  
<TEXT>/NN  
GENERAL/NNP ACCIDENT/NNP ,/, the/DT leading/VBG British/JJ insurer/NN ,/,  
said/VBD yesterday/NN that/DT insurance/NN  
claims/NNS arising/VBG from/IN Hurricane/NNP Andrew/NNP could/MD 'cost/VB  
it/PRP as/RB much/JJ as/IN Dollars/NNPS 40m.'/CD  
Lord/NNP Airlie/NNP ,/, the/DT chairman/NN who/WP was/VBD addressing/VBG  
an/DT extraordinary/JJ shareholders'/NN  
meeting/NN ,/, said/VBD :/: 'On/VBG the/DT basis/NN of/IN emerging/VBG  
information/NN ,/, General/NNP Accident/NNP  
advise/VB that/IN the/DT losses/NNS to/TO their/PRP\$ US/PRP operations/NNS  
arising/VBG from/IN Hurricane/NNP Andrew,/NNP  
which/WDT struck/VBD Florida/NNP and/CC Louisiana/NNP ,/, might/MD in/IN  
total/JJ reach/NN the/DT level/NN at/IN which/WDT  
external/JJ catastrophe/NN reinsurance/NN covers/VBZ would/MD become/VB  
exposed'./JJ  
What/WP this/DT means/VBZ is/VBZ that/IN GA/NNP is/VBZ able/JJ to/TO pass/VB  
on/IN its/PRP\$ losses/NNS to/TO external/JJ  
reinsurers/NNS once/RB a/DT certain/JJ claims/NNS threshold/NN has/VBZ  
been/VBN breached./JJ  
It/PRP believes/VBZ this/DT threshold/NN may/MD be/VB breached/VBN in/IN  
respect/NN of/IN Hurricane/NNP Andrew/NNP  
claims./CD  
However/RB ,/, if/IN this/DT happens/VBZ ,/, it/PRP would/MD suffer/VB a/DT  
post-tax/JJ loss/NN of/IN Dollars/NNPS 40m/CD  
(Pounds/NNS 20m)./CD  
Mr/NNP Nelson/NNP Robertson/NNP ,/, GA/NNP 's/POS chief/JJ general/JJ  
manager/NN ,/, explained/VBD later/RB that/DT the/DT  
company/NN has/VBZ a/DT 1/2/CD per/IN cent/NN share/NN of/IN the/DT  
Florida/NNP market./JJ It/PRP has/VBZ a/DT branch/NN in/IN  
Orlando./NNP  
The/DT company/NN 's/POS loss/NN adjusters/NNS are/VBP in/IN the/DT area/NN  
trying/VBG to/TO estimate/VB the/DT losses./CD  
Their/PRP\$ guess/VBP is/VBZ that/IN losses/NNS to/TO be/VB faced/VBN by/IN  
all/DT insurers/NNS may/MD total/VB more/JJR than/IN  
Dollars/NNS 8bn./CD  
Not/RB all/DT damaged/VBN property/NN in/IN the/DT area/NN is/VBZ  
insured/VBN and/CC there/EX have/VBP been/VBN  
estimates/NNS that/IN the/DT storm/NN caused/VBN more/JJR than/IN  
Dollars/NNPS 20bn/CD of/IN damage./JJ  
However/RB ,/, other/JJ insurers/NNS have/VBP estimated/VBN that/IN  
losses/NNS could/MD be/VB as/RB low/JJ as/IN

Dollars/NNS 1bn/CD in/IN total./JJ  
Mr/NNP Robertson/NNP said/VBD :/: 'No/UH one/CD knows/VBZ at/IN this/DT  
time/NN what/WP the/DT exact/JJ loss/NN is'./CD  
</TEXT>/NN  
<PUB>The/NNP Financial/NNP Times/NNP  
</PUB>/NN  
<PAGE>/NN  
London/NNP Page/NNP 16/CD  
</PAGE>/NN  
</DOC>/SYM