# Exporting Phrases: A Statistical Analysis of Topical Language

Amy M. Steier and Richard K. Belew
Cognitive Computer Science Research Group
Computer Science & Engr. Dept. (0114)
University of California - San Diego
La Jolla, CA 92093
steier@cs.ucsd.edu, rik@cs.ucsd.edu

## Abstract

*This paper describes preliminary experiments documenting significant variations in word usage patterns within topical sublanguages. As some phrases have very different collocational patterns than their constituent words, we look beyond occurrences of individual words, to consider word phrases. The mutual information statistic is used to measure the information content of phrases beyond that of their constituent words. We find that specialized topic areas give rise to phrases with very descriptive constituents which are then "exported" into general vocabulary. These phrases are also much more informative as word pairs* outside *the topic area than within it. Further, we find evidence of an intriguing "self-similar" regularity in this exporting relation across different hierarchical levels of topical areas.*

## 1 Introduction

The assumption is often made in information retrieval (IR) and corpus-based linguistics that the documents of interest are part of a single, homogeneous and unstructured collection. If the text corpus is of small or moderate size, topically well-focused and generated by one author or a small group of authors sharing a common vocabulary, this can be a reasonable and useful simplification. But as machine-readable corpora increase in size, it becomes more and more likely that significant variations in word usage patterns will be observed *within restricted subsets* of the collections.

We are particularly interested in variations arising from what might be viewed as technical sublanguages. In technical writings, for example scientific publications or legal opinions, it is often the case that specialized vocabularies (jargon, terms of art) evolve to describe aspects of the world that are too abstract or obscure to become part of common parlance. This paper describes some preliminary experiments we have performed documenting significant variations in word usage within specialized areas of the law (e.g., labor relations), as compared to a corpora covering all of U.S. case law.

In these experiments we look beyond occurrences of individual words, to consider word pairs or phrases. Simple phrases provide an attractive first step beyond the simple "bag of words" techniques, generally as-

sociated with IR. The focus of our experiments is to study how the informational content of phrases changes within topically restricted areas. We use the mutual information measure as the statistic best reflecting the informational content of these word pairs beyond that of their constituent words. Our central (albeit preliminary) finding is somewhat contrary to our initial expectations. We have been intrigued to find that a distinguishing characteristic of specialized topical areas is that they give rise to phrases that become much more informative as word pairs *outside* the area than within it. More specifically, if a phrase contains a word that standard automatic indexing metrics suggest is a good content descriptor within a topically restricted subset of documents, that phrase will consistently have a higher mutual information value outside that topic.

## 2 Identifying Phrases

Phrases are interesting in part because of the complex way in which meanings associated with constituent words are combined to form more elaborate semantic expressions. D.A. Cruse has defined a "transparent" expression to be one whose meaning is derived directly from that of its constituent words; and "opaque" phrases as those whose semantics cannot be attributed to the simple composition of its constituent terms [4]. Halliday has made similar distinctions between "simple," "compound" and "phrasal" lexical items [8]. Halliday writes that often a phrase can act very much like a lexical unit in and of itself, and that often a phrase can have very different collocational patterns than the sum of its parts. From a practical IR perspective, these linguistic issues become the question of just when indexing word compounds offers advantages over simpler indexing of individual words.

Like every other linguistic phenomena, "phrase" proves to be an extremely complicated construct, especially as it is taken to refer to syntactic relations that depend on elaborate grammatical systems. Because it forms a natural progression from IR's "bag of words" methodology, we follow other recent work that considers a minimalist notion of phrase, based on simple collocation [15][7][9]. As we are most interested in simple word compounds (such as "social security"), we use sequential word pairs, or bigrams to identify potentially interesting word co-occurrences. Church and Hanks have shown that compounds have a very fixed word order and that the average separation is one word [9]. Therefore, we found it reasonable to restrict ourselves to a window of one when parsing for bigrams. The dictionary gives one definition of a phrase as "a word or group of spoken words that is perceived momentarily as a meaningful unit and that is preceded and followed by pauses [16]." It is this notion of phrase as a "meaningful unit" that we adapt in this paper.

Wittgenstein's "meaning through use" philosophy provides another way of quantifying these notions [19]. Words have meanings derived from their use both independent of, and with respect to, use within a phrase. Conversely, the phrasal meaning is drawn from the meaning of the constituents as well as from the "use" of the phrase. As a phrase becomes more frequently used, we hypothesize that a phrase's meaning draws less on the experience of the constituents' uses in other contexts and more on the experience of the phrase in its particular context. [1] Mutual information therefore becomes a very natural measure of the dis-

[1] In other words, less meaning is drawn from the "transparent" semantics of the constituents, while more meaning becomes related to the direct experience of using the phrase.

parity between a phrase's use and the independent use of its constituent words.

## 3 The Use of Phrases in Information Retrieval

Amsler expounds on what he sees as the "Two Half-Truths of Information Science" that made text storage and retrieval a technique with hidden fatal flaws [1]:

1. A word is a contiguous sequence of alphabetic characters
2. Computers can retrieve information about a concept by retrieving occurrences of a word form which represented that concept

The first half-truth ignores the fact that often mult-word forms will act much like a unit. For example, nominal compounds (e.g. `social security`), phrasal verbs (e.g. `married off`), idioms and common phrases. The emphasis in theoretical linguistics has always been on the productive aspect of language, but in reality there are many phrases with non-compositional meaning as well (e.g. `hard liquor` or `funny bone`). The second half-truth ignores the fact that words and concepts are simply not in a one to one relation. There are many words with multiple meanings and many concepts that require multiple words to be expressed.

There have been many recent attempts to incorporate phrases into an indexing language [6] [11] [12] [14] [17]. This research has resulted in many new questions and issues on just how to extract and use phrases. For example, should we extract phrases syntactically or statistically? Of these extracted phrases, should we allow only the phrase itself as an index term, or should we allow it constituents to be index terms as well? For example, there are many opaque phrases such as `artificial intelligence` or `home run` where the meaning of the phrase is far from the typical meaning associated with its constituents. In these cases, although the phrase may be a useful descriptor or index term for a particular topic, the phrase's constituents are not.

There has recently been an increased interest in phrase-oriented lexicons [1] [5] [3] [10]. Interestingly enough, there are dictionaries that already contain a substantial number of phrases. Ahlswedee et al found one seventh of the entries in a machine-readable version of *Webster's Seventh New Collegiate Dictionary* to be phrases [13]. Many words, however, are often missing from the dictionary. Furthermore, it has been hypothesized that since phrases are even more dynamic than words, that this would especially be the case for phrases or "collocational expressions" [3]. Not only are collocational expressions context dependent but they are also transient in nature. New trendy expression (especially in the media) will come and go all the time.

A very interesting question is then how can dictionaries be automatically updated with important phrases, or as Choueka puts it, the goal is to automatically extract:

> *... interesting collocational expressions - i.e. sequences of words whose unambiguous meaning cannot be derived from that of their components, and which therefore require specific entries in the dictionary.*

Choueka's technique was to extract sequences of words which had a minimal frequency of 10, neither began or ended with a "function-word" [2], and contained none of

[2] Function-words are prepositions, pronouns, articles, conjunctions, question words, etc

the most frequent words [3]. He also made sure not to include sequences (such as `York Times`) which only ever occur as a part of some longer sequence (`New York Times`).

Choueka's results were only verified manually, but the phrases extracted did seem very useful. Some examples are: `white house`, `stock exchange`, `super bowl` and `middle east`. Choueka also hypothesized that it might be useful to try incorporating morphological and syntactic information and to:

> *Define somehow a "binding degree" of an expression as a measure of the degree with which the different words in the expression "attract each other".*

Choueka's work is extremely important because he is proposing a technique which will *automatically* extract from the corpora those phrases that *with respect to that corpora* act like lexical units. The point here is that just what are these lexico-phrasal units will change year to year, subject to subject and corpora to corpora. Any technique such as this that avoids a costly labor intensive manual effort is very pragmatic.

In the experiments described in this paper, we use mutual information as a way to measure the "binding degree" of an expression. We show that across the difference sub-contexts of a large corpus, the extent to which a phrase acts like a lexical unit can vary. This variation, however, turns out to be far from random. Our results show if a phrase has high mutual information with respect to the entire collection, then it will have a depressed mutual information value with respect to a specific topic area if and only if at least one of its constituents is very descriptive of that topic area.

[3] Such as numbers, dates, etc.

# 4 The Corpus

Over more than a hundred years, as a publisher of various court reporters that form the core of all legal libraries, the West Publishing Company has been continually developing a taxonomy covering all case law. This taxonomic classification system is used by West's editors to organize case digests, and is taught universally in law schools as a method of legal research [18]. The top level of this hierarchy has seven main divisions which are divided into subdivisions, and then into approximately 400 "topics." Each topic is divided further into "key numbers," with key number's often divided into divisions, subdivisions, etc. In short, the West Key number system is one of the richest hierarchical manual indexing systems in existence.

The data used in our experiments was taken from an experimental collection of approximately 12,000 Federal Court cases provided by West Publishing Company, covering virtually all topics of U.S. case law. The preliminary experiments reported here were restricted to "headnotes" associated with these cases. These headnotes are precis generated by West's editors to capture the central points of each judicial opinion. As an individual court case may involve many different points of law, there are often several headnotes associated with each case. Within this collection, there are almost 50,000 different headnotes. Collectively, these headnotes contain over three million words.

# 5 Methodology

Headnotes were grouped together by topic numbers to get a total of 339 "documents," each corresponding to a topic of law for which our collection contained representa-

tive cases. After some simple stemming, [4] we extract all bigrams from within each document. We then filtered out any bigram that crossed a sentence boundary or similar punctuation marker, as well as any bigrams containing "noise words". [5] Since the mutual information measure can become unstable when counts are very low, we also filtered out any bigram from within a topic file which occurred in the topic less than three times or outside the topic file less than three times.

We then compute, for each bigram within a topic file, a mutual information value with respect to the topic file, $MI_t$, as well as a value with respect to the entire collection of topic files, $MI_c$. The mutual information measure is:

$$MI(w_1, w_2) = log \frac{Prob(w_1, w_2)}{Prob(w_1)Prob(w_2)}$$

Here, $Prob(w_j)$ is the frequency of $w_j$ divided by $N_W$, the size of the corpus. $Prob(w_1, w_2)$ is the frequency of bigram $(w_1, w_2)$ divided by $N_W$. In computing $MI_t$, $N_W$ is the number of words in that particular topic file. In computing $MI_c$, $N_W$ is the number of words in all topic files. The consequence of this measure is if the constituents of a phrase occur together much more often than chance, the phrase will have a high mutual information value. The higher the mutual information value is between a pair of words, the more informative that pair is as a phrase. It is not that the phrase is necessarily more contentful but that our interpretation of the phrase is less easily derived from the typical meaning associated with its constituents.

Next, each individual *word* within a topic file is given an index term weight based on a variant of the term frequency $\times$ inverse document frequency weighting scheme devised by Salton and Buckley [2]. If $F_{ij}$ represents the frequency of term $j$ in document $i$, $DF_j$, the document frequency of term $j$, $N_D$, the total number of documents, and $N_i$, the number of words in document i, then $TW_{ij}$, the term weight of term $j$ in document $i$, is given by the following formula:

$$TW_{ij} = \frac{F_{ij} \times \log(N_D/DF_j)}{\sqrt{\sum_{k=1}^{N_i}(F_{ik} \times \log(N_D/DF_k))^2}}$$

## 6 Data Analysis

Our first analysis was to compare $MI_t$, the mutual information values within a topic, to $MI_c$, the mutual information with respect to the entire collection. Initially, we found much variation in the mutual information, but it was difficult to see any patterns. There seemed to be just as many cases where $MI_c$ exceeded $MI_t$ as there were cases where $MI_t$ exceeded $MI_c$. Figure 1 graphs this relationship for topic 232A, Labor Relations. Since we were surprised by the cases where $MI_c$ exceeded $MI_t$, we decided to restrict our analysis to those phrases with a high informational content in the collection as a whole. We took $MI_c$ greater than or equal to six to mean a high phrasal information content. Note that even if you only consider phrases with this restriction, there are still just as many cases where $MI_c$ exceeds $MI_t$ as there are vice versa.

Our next step was then an analysis of those phrases with large mutual information differences. Table 1 shows an example set of phrases where $MI_t$ exceeds $MIc$ and Table 2 shows some examples where $MI_c$ exceeds $MI_t$.

In studying these phrases, what stands out is how apropos to the Labor Relations

[4] Our stemming consisted only of converting all plural nouns to their singular form.

[5] For our purposes, a noise word is a non-content word such as an article or preposition, as well as legal abbreviations of statute sections.

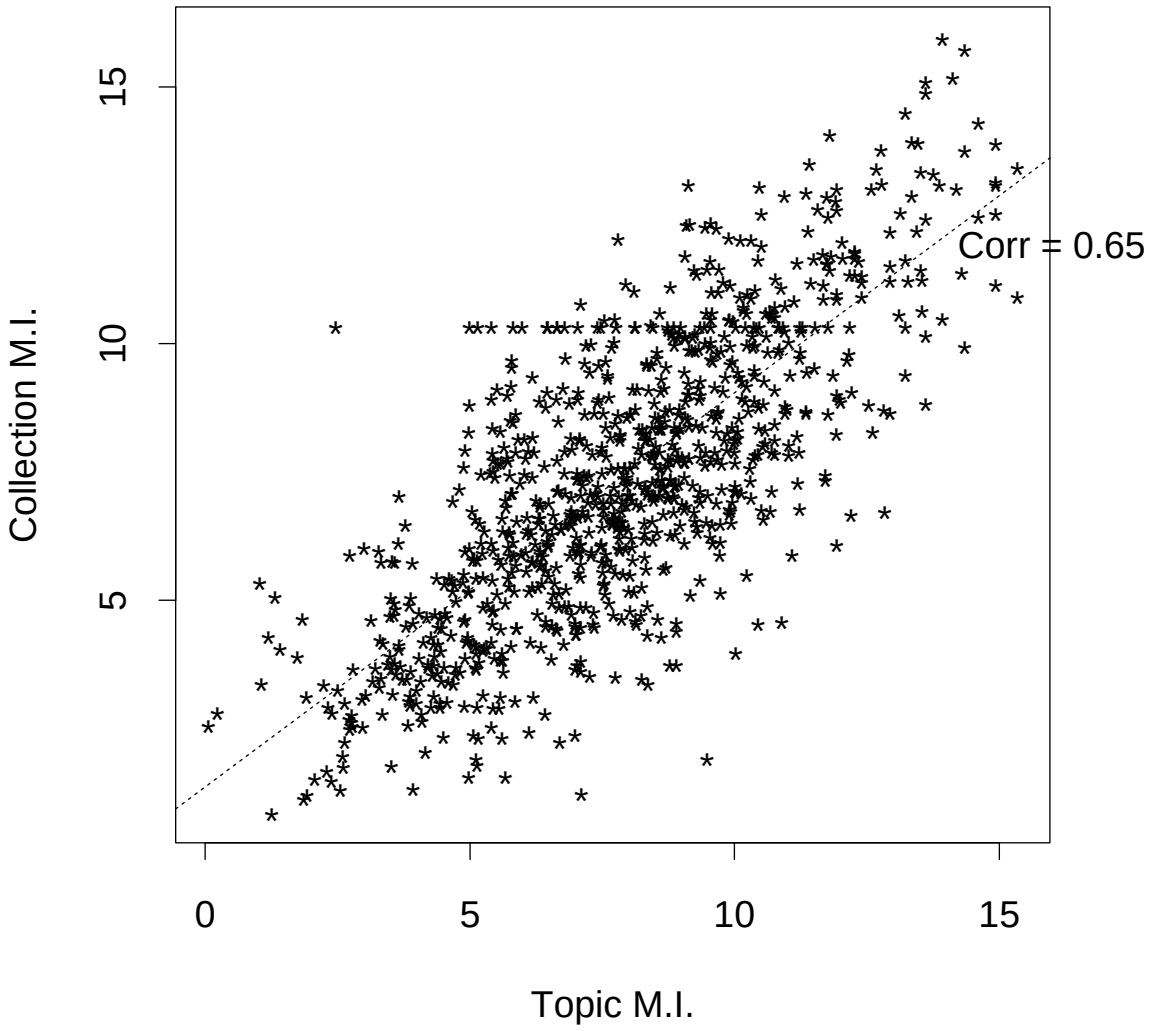


**Figure 1: A comparison of mutual information (M.I.) for the topic Labor Relations. The dotted line in the graph maps the least squared fit. Correlation is shown as 0.65.**

| *PHRASE* | $TW_M$ | $MI_t$ | $MI_c$ |
|---|---|---|---|
| JUST CAUSE | .015 | 10.39 | 7.87 |
| APPLICABLE REGULATION | .009 | 9.73 | 6.51 |
| PUBLIC INTEREST | .010 | 10.01 | 7.71 |
| APPROPRIATE FORUM | .007 | 11.71 | 7.49 |
| FRAUDULENT INTENT | .005 | 12.54 | 8.84 |
| CONSTRUCTIVE KNOWLEDGE | .004 | 13.61 | 10.17 |
| DISCRETIONARY AUTHORITY | .014 | 11.18 | 8.23 |
| GOOD STANDING | .013 | 10.01 | 7.26 |
| REGULAR RATE | .012 | 10.37 | 8.92 |
| BEST INTEREST | .010 | 10.58 | 8.35 |

**Table 1: Example phrases where the mutual information within the topic Labor Relations ($MI_t$) exceeds that within the collection ($MI_c$.)**

| *PHRASE* | $TW_M$ | $MI_t$ | $MI_c$ |
|---|---|---|---|
| MINIMUM WAGE | .057 | 9.48 | 11.50 |
| WORKER COMPENSATION | .043 | 5.98 | 10.37 |
| COLLECTIVE BARGAINING | .185 | 7.80 | 12.08 |
| UNION MEMBER | .449 | 6.31 | 8.91 |
| OCCUPATIONAL SAFETY | .050 | 10.51 | 12.56 |
| OVERTIME PAY | .062 | 5.40 | 8.95 |
| LOCAL UNION | .449 | 6.03 | 8.17 |
| PENSION PLAN | .076 | 8.42 | 10.40 |
| LABOR RELATION | .495 | 4.99 | 8.84 |
| LABOR STANDARD | .495 | 5.78 | 9.20 |

**Table 2: Example phrases where the mutual information within the collection ($MI_c$) exceeds that within the topic Labor Relations ($MI_t$.)**

topic the phrases in Table 2 are, whereas this is not the case in Table 1. $TW_M$ (maximum term weight) is the larger of the two term weights associated with the phrases constituents with respect to the topic Labor Relations. All term weights range between 0 and 1, and the median term weight for a topic is about .03. Note that there is a strong tendency for phrases with high $TW_M$ to have a depressed mutual information value within the topic.

Figure 2 shows the relationship between maximum term weight and the difference in mutual information ($MI_t$ - $MI_c$) for phrases where $MI_c >= 6$. Since our term weighting algorithm assigns to the majority of terms a very low weight, and then progressively less terms a higher and higher weight, we use the log of the maximum term weight to more aptly show its relationship to the difference in mutual information. The correlation of the two parameters in this graph is -.62, which we found very typical across the other topics within the collection. That is, while bigram depression is defined with respect to a particular topical area, the general phenomena of depression is observed in every topical area: across the entire collec-

tion, 93% of all bigrams with high maximal term weight (0.1 or greater) had depressed mutual information values.

The conclusion we draw from this phenomenon is that the more descriptive the constituents of a phrase are with respect to a specific topic, the more transparent that phrase is within the topic and the more opaque it is outside the topic. Take, for example, the phrase PENSION PLAN. This phrase seems to be much more opaque (i.e. have a much higher phrasal information content) across the entire collection. Yet within the Labor Relations topic, the constituents PENSION and PLAN are very descriptive of the central issues. Not only do these words occur together as PENSION PLAN, but they often occur separately (e.g., PENSION FUND, PENSION BENEFIT, PENSION BOARD, INSURANCE PLAN, WELFARE PLAN.)

Our next step was to do a more in depth study of three different topic areas. For each one, we re-divided up the headnotes into files corresponding to sub-topics. We then replicated our original experiments, but this time comparing $MI_{st}$, the mutual information within a sub-topic, to $MI_t$, the mutual information with respect to the entire topic. We were intrigued to find that our results at the sub-topic level duplicated those at the topic level.

Once again, we found that there was a strong tendency for phrases with high term weights to have a depressed mutual information value within the sub-topic. Within the Labor Relations sub-topic of Availability of Judicial Remedies, we found that the correlation between maximum term weight and difference in mutual information to be -.51 which is typical of the other sub-topics within Labor Relations. Table 3 depicts an example set of phrases where $MI_t$ exceeds $MI_{st}$ within this particular Labor Relations sub-topic.

A good example is that of GRIEVANCE PROCEDURE. At the topic level, this phrase had a maximum term weight of .11 and an $MI_c$ value of 9.14. Although the phrase's constituents were descriptive of the central issues in Labor Relations, they are even more descriptive of the central issues in the sub-topic of Availability of Judicial Remedies. Similarly, although this phrase was more transparent at the topic level than it was with respect to the collection as a whole, it is even more transparent at the sub-topic level. In studying the other bigrams that occur within this sub-topic, we found that although the constituents GRIEVANCE and PROCEDURE do co-occur together within the sub-topic, they also occur very often separately (e.g., GRIEVANCE REMEDIES, GRIEVANCE PROVISIONS, GRIEVANCE MECHANISMS, SETTLEMENT PROCEDURES, REMEDIAL PROCEDURES, ARBITRATION PROCEDURES).

## 7 Interpretation

Our initial expectation was that, for many technical phrases, the informational content of the phrase would be much higher within a topical area (semantically associated to the phrase) than across the entire legal corpus. However, our results showed that many phrases exhibited the opposite behavior. In fact, there were as many phrases where mutual information was decreased within a topic as there were phrases where mutual information increased within a topic. Our most central finding, however, is that when you consider only phrases with high ($>= 6$) mutual information with respect to the entire collection, a decrease of mutual information within a topic area turns out to be highly correlated to just how descriptive the constituents of that specific topic area are. In other words, bigrams with high informational content across the entire

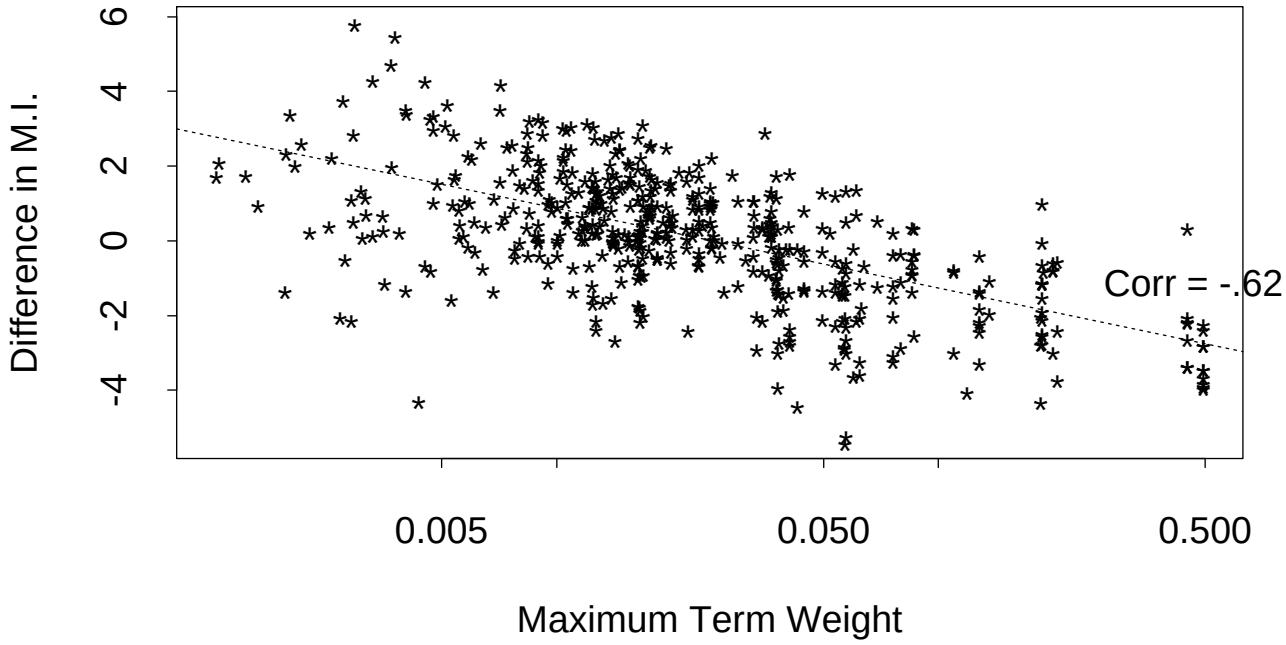


**Figure 2: Log of the Maximum Term Weight to difference in $MI$ ($MI_T$ - $MI_C$) for the topic Labor Relations. Only phrases with $MI_C$ >= 6 are used. The dotted line in the graph maps the least squared fit.**

collection have *depressed* informational content (i.e., convey less information) — as bigrams — within their semantically-related topical area *if, and only if,* the bigrams constituents are good descriptors of that topic area.

Our interpretation of these results is that technical phrases (or more narrowly, bigrams) are often "deconstructed" into their constituents within a particular topical sublanguage. While `PENSION PLAN` works extremely well as an informational unit across the entire legal domain, the constituent concepts of `PENSION` and `PLAN` are often examined in detail and independently in the topical area dealing with this aspect of the law. The semantic nuances that are explored in detail within a topical area are then left behind as this phrase is "exported" into general vocabulary, where dominant use of the constituent words are as part of the phrase.

Further, we find evidence of an intriguing "self-similar" regularity in this exporting relation, both from one of West's topical areas to the general corpus, and from within further specialized sub-topics to the relatively more general topic. This leads us to conclude that "opacity" is a relative concept. An opaque phrase whose mean-

| *PHRASE* | $TW_M$ | $MI_{st}$ | $MI_t$ |
|---|---|---|---|
| GRIEVANCE PROCEDURE | .270 | 6.54 | 8.38 |
| INTRAUNION REMEDY | .144 | 7.72 | 10.11 |
| BARGAINING AGREEMENT | .181 | 6.46 | 7.29 |
| MEDIATION BOARD | .120 | 7.51 | 8.46 |
| ADMINISTRATIVE REMEDY | .144 | 7.64 | 8.60 |
| ADJUSTMENT BOARD | .120 | 7.43 | 8.43 |
| COLLECTIVE BARGAINING | .161 | 6.71 | 7.80 |
| CONTRACTUAL REMEDY | .144 | 6.40 | 7.71 |
| GRIEVANCE MACHINERY | .270 | 6.80 | 8.70 |

**Table 3: Example phrases where the mutual information within the topic Labor Relations ($MI_t$) exceeds that within the sub-topic Availability of Judicial Remedies ($MI_{st}$.)**

ing appears to have little relation to that of its constituent words may in fact be transparent within some restricted sublanguage, perhaps associated with its genesis. Of course, this sublanguage may no longer exist, or it may simply require closer analysis of restricted portions of the general textual corpus.

## 8 Conclusion

We consider these conclusions to be preliminary and little more than working hypotheses. More recent experiments suggest that the results we've discussed in this paper are dependent upon the mutual information statistic. Repeating these experiments using another statistical measure (chi-square), our original "export" relation still seems to exist, but the relationship is no longer symmetrical. That is, if a bigram's constituents are good descriptors of a topic area, that bigram will consistently have a depressed informational content within that topical area but the converse is not necessarily true. We attribute this change to a particular characteristic of mutual information, viz. converging to $\frac{N_W}{F(w_1,w_2)}$ when a bigram's constituents are perfectly correlated.[6] Therefore, a highly correlated word pair that is infrequent, will have a higher mutual information than one that is more frequent. We are still, however, studying the results of these more recent experiments.

In our future research, we hope to gain more insight into the similarity and differences between mutual information and some of the other statistical metrics commonly used to measure association between word pairs. We also plan to study collocational patterns in the judicial opinions themselves (vs. the headnote text), in other non-legal textual corpora that may also allow the formation of technical sublanguages (e.g., scientific writings), and ultimately in non-technical natural language. More immediately, even these preliminary results seem to suggest that IR would do well to consider topical organizations within a large corpus. Just as keywords that make good descriptors for a large collection may not be useful discriminators for a subcollection, phrases too seem to be highly context-dependent.

What if the topical structure of a collec-

[6] Here, $N_W$ is total number of words in the collection and $F(w_1, w_2)$ is the frequency of the bigram.

tion is not known *a priori*? We've shown in this paper that the mutual information of a phrase with respect to the entire collection cannot be reliably used within all topic areas. We have not yet compared the weights of the phrases and words in the headnotes of a topic area to those associated with individual opinions; our most immediate research goal is to establish whether or not this correlation does exist. If so, we hypothesize that one might be able to use an individual documents term weights to signify a depression of informational content within a phrase, thereby providing clues to subject topicality.

## Acknowledgment

The authors wish to thank the West Publishing Company for the use of this tremendously rich collection of documents, specifically Howard R. Turtle. We also wish to thank Apple (Advance Technology Group) for their financial support, specifically Dan Rose.

## References

[1] R.A. Amsler. Research toward the development of a lexical knowledge base for natural language processing. In *Proceedings of the 12th International Conference on Research and Development in Information Retrieval*, pages 242–249, Cambridge, Mass., June 25-28 1989.

[2] G. Salton C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[3] Y. Choueka. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of RIAO '88; User-Oriented content-based text and image handling*, pages 609–623, Cambridge, Mass., 1988.

[4] D.A. Cruse. *Lexical Semantics*. Cambridge University Press, New York, 1986.

[5] U. Zernik M.G. Dyer. The self-extending phrasal lexicon. *Computational Linguistics*, 13(3-4):308–327, July-December 1987.

[6] J. Fagan. *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Nonsyntactic Methods*. PhD thesis, Cornell University, January 1988.

[7] J. Fagan. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40(2):115–132, 1989.

[8] M.A.K. Halliday. *Lexis as a Linguistic Level*. Longmans, London, 1966.

[9] K. Church P. Hanks. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1), March 1990.

[10] D.J. Besemer P.S. Jacobs. Flush: A flexible lexicon design. In *25th Annual Meeting of the Association for Computational Linguistics*, 1987.

[11] D.D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th International Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, June 21-24 1992.

[12] D.A. Evans K. Ginther-Webster M. Hart R.G. Lefferts I.A. Monarch. Automatic indexing using selective nlp and first-order thesauri. In *Proceedings of RIAO '91; User-Oriented content-based text and image handling*, pages 624–643, Barcelona, Spain, 1991.

[13] T. Ahlswede J. Anderson M. Evens S.M. Li J. Neises S. Pin-Ngern. Automatic construction of a phrasal thesaurus for an information retrieval system from a machine readable dictionary. In *Proceedings of RIAO '88; User-Oriented content-based text and image handling*, pages 597–608, Cambridge, Mass., 1988.

[14] L.P. Jones E.W. Gassie Jr. S. Radhakrishnan. Index: The statistical basis for an automatic conceptual phrase-indexing system. *Journal of the American Society for Information Science*, 41(2):87–97, 1990.

[15] Y. Maarek F. Smadja. Full text indexing based on lexical relations. In *Proceedings of the 12th International Conference on Research and Development in Information Retrieval*, Cambridge, MA, June 1989.

[16] J. Stein L. C. Hauck P.Y. Su, editor. *The Random House College Dictionary.* Random House Inc., New York, revised edition, 1980.

[17] T. Strzalkowski B. Vauthey. Information retrieval using robust natural language processing. In *30th Annual Meeting of the Association for Computational Linguistics*, 1992.

[18] *WESTLAW Reference Manual.* West Publishing Company, St. Paul, MN, 3rd edition, 1989.

[19] L. Wittgenstein. *Philosophical Investigations.* Macmillan Publishing Co., Inc., New York, third edition, 1958. Translated by G. E. M. Anscombe.