A Comparison Between Statistically and Syntactically Generated Term Phrases

Gerard Salton
Chris Buckey*

TR 89-1027
July 1989

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501

*This study was supported in part by the National Science Foundation under grant IRI-87-02735

A Comparison Between Statistically and Syntactically Generated Term Phrases

Gerard Salton and Chris Buckley*

Abstract

It is customary to use single terms (words) or terms in context (phrases) as indexing units for the representation of natural-language text content. There is evidence that term phrases may provide some advantages over the use of single terms for text content representation. This note presents an evaluation of the expected usefulness of automatic term phrase generation systems involving syntactic processing compared with methods based only on the statistical co-occurrence characteristics between individual text words.

1. Statistical and Syntactic Phrases

The idea of using statistical associations between words for the generation of text indexing units goes back at least to the early 1960s. [1-3] It was suggested at that time that sets of words occurring sufficiently often in specified contexts in a document collection might be used to form meaningful indexing units for the representation of text content. A typical prescription for the generation of such "statistical phrases" states, for example, that an indexing phrase would be formed "when all **n** components of the phrase have been found to occur at least **m** times in the sentences of a document". [4]

More generally the following criteria are usable for the formation of statistical phrases:

a) the term weight of the individual phrase components;

b) the frequency with which the phrase components co-occur in the document texts;

c) the contexts within which the phrase components co-occur, measured, for example, by the number of intervening text words between phrase components.

There is some evidence that phrases generated by using word co-occurrence characteristics are valid only locally for the collections form which the phrases were derived, rather than globally for all text environments. [5]

*This study was supported in part by the National Science Foundation under grant IRI-87-02735.

Nevertheless, recent exhaustive tests of statistical phrase generation systems have shown once again that the use of statistical phrases for content identification can improve the recall-precision performance of text retrieval by up to 20 percent compared with the use of single-term indexing units. [6]

Although statistical phrase units have been used successfully in various text processing applications, it has been claimed many times that a text identification system based on statistical word associations cannot represent the last word, because text content is not determinable simply by the use of statistically, or probabilistically associated text words. For these reasons, more refined linguistic analysis methods have been proposed for the generation of text indexing units. The preferred *semantic* analysis systems based on prestored knowledge representing the structure of particular topics or disciplines are not available when the document collections cover arbitrary subject matter, because the relevant knowledge structures are not easily captured in that case. However, the existing *syntactic* analysis systems are sufficiently versatile to be able to handle structural analyses of arbitrary samples of natural language text. Syntactic analysis systems have, in fact, previously been used in text retrieval environments. [7-11] For this reason, a new attempt is made here to compare the expected usefulness of statistically and syntactically produced term phrases.

The experiments which follow are based on a statistical analysis of the full text of a recent volume on Text Processing. [12] The syntactic phrase evaluation covers two particular chapters of that book (chapters 5 and 6) dealing with text compression and text encryption, respectively, and comprising about 12,000 words of text. For analysis purposes, the text of the book was subdivided into 1,400 local documents, each approximately one paragraph in length. Automatic indexing methods were then used to assign indexing units separately to each local document, and the evaluation was based on an assessment of the usefulness of the phrase indexing units for local document representation.

The experimental design and test results are described in the remaining sections of this note.

2. Statistical Phrase Evaluation

A standard automatic indexing techniques consists in distinguishing the individual text words, deleting entries found on a special list of stop words (such as "and", "or", "but", "of", etc.), reducing the remaining words to word stem form by appropriate suffix deletion, and finally assigning a term weight to the remaining word stems. [13] A

useful term weighting system, known as the (tf × idf) weighting system, takes into account the frequency of a term (word stem) in a given document, the number of documents to which a term is assigned, and the document length (that is, the number of terms assigned to a given document). [14,15] Typically, a good term is one with a high internal frequency in a document, but a low overall collection frequency. Such a term is then assigned to a document for content representation.

Fig. 1 shows the text of local documents 1 and 2, representing the beginning of chapter 5 of [12]. The ten most highly weighted terms (word stems) according to the (tf × idf) weighting system for the two sample documents are shown in Table 1(a). In each case, the table shows the document number, the concept number assigned to each term during processing (column 2 of the table), the term weight (column 3) and finally the corresponding word stem (column 4). In the (tf × idf) system, the term weights vary between 0 and 1. It may be noted that even the best terms for the two sample documents ("text" and "occur", respectively), receive a relatively low weight.

The statistical phrases assigned to the sample documents are shown in Table 1(b). These phrases are generated by using collections of word stems occurring adjacently in the respective document texts, such that all phrase components appear among the ten most highly weighted stems for each document. Such a strategy produces 7 phrases for document 1 and 11 phrases for document 2. If the phrase lists are further restricted to those exhibiting a phrase frequency of at least 2 in each document, the number of assigned phrases shrinks to 2 and 3 for the two documents, including "storage space", "text compression", "frequent word", "letter occurrence", and "word occurrence", respectively. The corresponding phrase entries are shown in Table 1(c) where the phrase frequency is given in parentheses. Phrases that are clearly in error are marked with an (x) in Table 1(b). These represent associated word pairs that do not appear to be semantically meaningful.

The statistical phrase generation system described here, consisting of highly weighted term sets occurring adjacently in document texts, was evaluated by considering each generated phrase, and rejecting those that did not appear to be semantically meaningful. Specifically a *phrase precision* factor was computed by using the proportion of semantically acceptable phrases generated for each of the 43 local documents of chapter 5, and the 50 local documents of chapter 6, respectively. For each document the phrase precision is then defined as

$$\text{Phrase Precision} = \frac{\text{Number of acceptable statistical phrases for a document}}{\text{Total number of generated statistical phrases for that document}}.$$

The average phrase precision figures for the documents of chapters 5 and 6 are shown in Table 2.

The top part of Table 2 shows that out of 534 and 334 phrase occurrences in the two sample chapters, 439 and 284 were judged to be semantically acceptable, producing a phrase precision of 0.822 and 0.850, respectively. The precision reaches 85 percent and 90 percent for the two chapters when the total number of phrases is reduced to 359 and 176 by applying a frequency threshold of 1, that is, by rejecting phrases found to occur only once in a given local document (see bottom of Table 2).

An analogous procedure can be used to obtain *phrase-recall* figures. For that purpose, each local document was examined, and a set of relevant content phrases was chosen manually representing the content of each particular document. The phrase-recall was then defied as the proportion of the manually assigned content phrases that was also contained in the list of automatically generated statistical phrases. Specifically, for each document, the phrase-recall is defined as

$$\text{Phrase Recall} = \frac{\text{Number of statistical phrases contained in the manually chosen relevant phrase set}}{\text{Total number of manually assigned relevant phrases for that document}}$$

The average statistical phrase recall for the local documents of chapters 5 and 6 is shown in Table 3.

As the table shows, the phrase recall figures do not reach the impressive levels produced earlier for the phrase precision. For the 500 phrases of chapter 5, the phrase-recall reaches about 55 percent, but for chapter 6 the recall is just below 40 percent. It is unclear to what extent the presence or absence of phrases deemed most appropriate by a human indexer will actually affect the recall performance of a retrieval system.

3. Syntactic Phrase Evaluation

Syntactic information can be used for the identification of term phrases in two main ways. First, a dictionary look-up program may be used to assigned syntactic tags (such as "noun", "adjective", "transitive verb", "preposition", etc.) to the text words, and adjacent word strings with particular tag sequences may then be recognized as proper noun phrases. Typically, all word sequences carrying noun and adjective tags could be used as indexing phrases. Such a strategy was used in certain earlier syntax-based indexing experiments. [7,8]

Alternatively, a full syntactic analysis system may be used to produce syntactic parse trees for the sentences of a document, and a syntactic phrase may then be defined as a subtree of the full parse tree exhibiting particular

structural characteristics. For the current experiments, the latter strategy was used. The full text of chapters 5 and 6 of [12] was analyzed using the PEG grammar (PLNLP English Grammar) developed at the IBM Research Laboratory for incorporation into the CRITIQUE text critiquing system. [16,17] A potential syntactic phrase was then recognized as consisting of a head-noun together with certain specified dependent structures located in the same subtree as the head-noun. [18,19]

The unmodified syntactic units derived from the parse trees for the sample documents of Fig. 1 are shown in Fig. 2. The entries of Fig. 2 consist of noun phrases, or single nouns without modifiers, followed by a list of markers identifying each phrase type as well as the type of grammar used in deriving the phrase. In general, the R-marker gives the phrase generation rule used to produce the phrase (R:1 for noun sequences, R:2 for adjective-noun sequences, etc.); the A-marker designates the type of article preceding the phrase (definitive, indefinitive, or no article at all); the P-marker at the end of the string specifies the grammar used to parse the corresponding sentence (P-1 for the regular grammar, and P-0 for an extended grammar with relaxed rules that is invoked when a parse tree is not obtainable with the regular grammar). The terminal sequence of digits designates the paragraph number and sentence number in the text where each phrase is located. [19]

The list of syntactic phrases contained in Fig. 2 contains some apparently appropriate indexing units such as "text compression" and "computer manipulation" for document 1. In addition, there are questionable entries such as "converting natural-language text representations", or "using document abstracts". The false entries are due to errors in the syntactic analysis process, or in the phrase generation system. [20] For the indicated sample phrases, "converting" and "using" were recognized as adjectives, rather than verbs, by the syntactic analysis system, resulting in the production of the corresponding false phrase constructs.

Table 4 presents the phrase-precision computations for the unedited syntactic phrase output for chapters 5 and 6. The table indicates that a substantially larger number of indexing units are obtained using the unedited syntax than for the statistical process previously given in Table 2. This is true especially because of the large number of single-term indexing units obtained by the syntactic method in addition to the multi-word sequences. Furthermore, the phrase-precision (60 and 74 percent for the phrases, and only 22 and 33 percent for the single terms) is lower than the precision of 80 to 90 percent obtained for the statistical output. It is obvious that the unedited syntactic output must be further edited and restricted if adequate syntactic output is wanted.

Two main strategies suggest themselves in order to obtain enhanced accuracy in the syntactic output. On the one hand, specific syntactic restrictions can be introduced by accepting only certain narrowly defined portions of the unedited syntactic output. Table 5 shows the reduced phrase set obtained by restricting the raw output of Fig. 2 to constructions identified by R1-P1 markers (a sequence of several nouns), R1-S1 markers (noun phrases extracted from titles and section headings), and R2-P1 markers (adjective-noun combinations), all constructions being generated by the regular (rather than the extended) grammar.

The output of Table 5 shows that the phrases obtained with these restricted definitions are mostly meaningful, the exceptions being marked with an (x). Phrases with a frequency indicator (such as (2)) appear more than once in the output. The phrase-precision for the phrases with syntactic tag restrictions is shown in Table 6. When only noun sequences or title phrases are accepted (R1-P1 or R1-S1 markers), a quite high phrase-precision of 92 percent is obtained as shown at the top of Table 6. When adjective-noun sequences (R2-P1 markers) are added, the phrase-precision for the resulting larger sets of 297 and 325 phrases, drops to 87 and 88 percent. These last precision figures are very close to the precision obtained for the frequency-restricted statistical phrases of Table 2.

An alternative way of restricting the raw syntactic output, and hopefully obtaining enhanced phrase precision figures, is to apply term weight restrictions to the phrase components. Table 7 shows the phrases obtained for sample documents 1 and 2 by choosing phrases containing highly weighted components. For the phrases in the left-hand column, *all* phrase components are among the ten most-highly weighted terms in the respective documents (see Table 1); the phrases on the right side of Table 7 contain at least one component among the ten most-highly weighted terms. The listing of Table 7 shows that once again very few unacceptable phrases (x) are obtained when term weight restrictions are applied to the syntactic output.

The term precision computations for the syntactic phrases with term weights restrictions appear in Table 8. When all phrase components are highly-weighted, 221 and 174 phrases are obtained for the two sample chapters, and the phrase precision of 80 and 89 percent is once again comparable to that obtained earlier for the statistical phrases of Table 2, and for the syntactic phrases with R1-P1 and R2-P1 tag assignments of Table 6. When the term weight restrictions are relaxed by accepting phrases with only one highly weighted component, the phrase-precision drops as shown in the lower half of Table 8.

To complement the phrase-precision output for the syntactic phrases (Tables 6 and 8), phrase-recall data can again be obtained by using the proportion of the manually selected relevant phrases covered by the various types of syntactic output. The relevant data are contained in Table 9. When severe restrictions are imposed on the syntactic output, and only a small number of phrases is generated, as in the output covering phrases with R1-P1 markers (see top of Table 9), the phrase recall is 40 percent and 27 percent for the two sample chapters, respectively. The recall increases when larger numbers of syntactic phrases are considered. For the phrases produced by the syntactic output with simple weight restrictions, the phrase recall reaches 62 percent and 54 percent as shown at the bottom of Table 9.

4. Summary

A summary of the phrase-precision results appears in Table 10. It is seen that the proportion of semantically acceptable phrases is very high (between 80 and 90 percent) for the frequency-restricted statistical phrases where all statistical phrases with internal document frequency greater than 1 are accepted. The syntactic process based on phrases with tag sequences of R1-P1 and R2-P1 (noun sequences, and adjective-noun sequences) generates a phrase set comparable in size to that of the frequency-restricted statistical method with comparable phrase-precision figures. The same is true of the weight-restricted syntactic phrases previously examined at the top of Table 8. When only a small number of phrases are needed, the phrase set specified by R1-P1 tags alone may be used; a slightly higher phrase precision (92 percent) may then be obtainable, as shown in the middle of Table 10.

The overall conclusion must be that the proportion of acceptable indexing phrases obtained by statistical and syntactic methodologies is approximately the same. This is true even though the number of phrases that are produced in common by the statisical and syntactic methods may be relatively small. The data at the bottom of Table 10 show that for the two sample book chapters, about 12 percent of the phrases appear in common in both syntactic and statistical outputs. Since the use of syntactic methodologies is much more onerous in terms of computer power, storage space, and program availability than the statistical methods, one finds that the syntactic methodology is not recommended at the present time for automatic term phrase generation systems.

A comparison of the phrase-recall figures contained in Tables 3 and 9 for statistical and syntactic phrases, respectively, shows that the recall output is somewhat higher for the syntactic output than the statisical. For exam-

ple, 359 phrases are obtained for chapter 5 using frequency-restricted statistical output with a phrase-recall of 0.401 (see Table 3). The comparable output in Table 9 for syntactic phrases with at least one highly-weighted component is 397 phrases and a phrase-recall of 0.626. For chapter 6 output sets of similar size are obtained for the frequency restricted statistical process (176 phrases), and the restricted syntactic system using only phrase component from the ten most highly weighted terms in each document (174 phrases). The phrase-recall again favors the syntactic process by a small amount (0.200 for the statistical phrases and 0.326 for the syntactic approach).

All the phrase recall figures are relatively low; moreover, it is not clear that a failure to reproduce automatically a particular set of manually chosen content phrases will impair the effectiveness of operational term indexing systems. In general, one may expect that the retrieval effectiveness of a set of indexing phrases depends chiefly on the phrase precision, that is, on the semantic viability of the phrases actually generated, rather than on the phrase-recall that is based on the emulation of an idealized phrase set. Given the equivalent phrase precision for statistical and syntactic processes, the advantage must rest with the simpler, statistical methodology.

References

[1] K. Sparck Jones, *Experiments in Semantic Classification*, Mech. Translation, 8:3-4, October 1965.

[2] L.B. Doyle, Indexing and Abstracting by Association, *American Documentation*, 13:4, October 1962, 378-390.

[3] V.E. Giuliano and P.E. Jones, *Linear Association Information Retrieval*, in *Vistas in Information Handling*, P. Howerton, editor, Spartan Books, Rochelle Park, NJ, 1963.

[4] G. Salton, *Automatic Information Organization and Retrieval*, McGraw Hill Book Company, New York, NY, 1968, p. 441.

[5] M.E. Lesk, Word-word Association in Document Retrieval Systems, *American Documentation*, 20:1, January 1969, 27-38.

[6] J.Fagan, The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval, *Journal of the ASIS*, 40:2, March 1989, 115-132.

[7] A.H. Klingbiel, Machine Aided Indexing of Technical Literature, *Information Storage and Retrieval*, 9:2, 79-84, and 9:9, 477-484, 1973.

[8] M. Dillon and A.S. Gray, Fully Automatic Syntactically Based Indexing System, *Journal of the ASIS*, 34:2, March 1983, 99-108.

[9] F.J. Damerau, Automatic Parsing for Content Analysis, *Communications of the ACM*, 13:6, June 1970, 356-360.

[10] D.J. Hillman and A.J. Kasarda, The Leader Retrieval System, *AFIPS Proceedings*, AFIPS Press, Montvale, NJ, 34, 1969, 447-449.

[11] A.F. Smeaton and C.J. van Rijsbergen, Experiments on Incorporating Syntactic Processing of User Queries into a Document Retrieval Strategy, *Proceedings of 11th International Conference on Research and Development in Information Retrieval*,, Y. Chiaramella, editor, Grenoble, France, June 1988, 31-52.

[12] G. Salton, *Automatic Text Processing*, Addison Wesley Publishing Company, Reading, MA, 1989.

[13] G. Salton, A Blueprint for Automatic Indexing, *ACM SIGIR Forum*, 16:2, Fall 1981, 22-38.

[14] K. Sparck Jones, A Statistical Interpretation of Term Specificity and its Application in Retrieval, *Journal of Documentation*, 28:1, March 1972, 11-21.

[15] G. Salton and C.S. Yang, On the Specification of Term Values in Automatic Indexing, *Journal of Documentation* 29:4, December 1973, 351-372.

[16] G.E. Heidorn, K.Jensen, L.A. Miller, F.J. Byrd, and M.S. Chodorow, The EPISTLE Text Critiquing System, *IBM Systems Journal*, 21:3, 1982, 305-326.

[17] K. Jensen, G.E. Heidorn, L.A. Miller and Y. Ravin, Parse Fitting and Prose Fitting: Getting Hold of Ill Formedness, *American Journal of Computational Linguistics*, 9:3-4, July-December 1983, 147-160.

[18] J. Fagan, Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Nonsyntactic Methods, Doctoral Dissertation, Cornell University, Technical Report TR 87-868, Department of Computer Science, Ithaca, NY, September 1987.

[19] J. Kaye, Back-of-the-Book Indexing Project, Term Report CS 490, Department of Computer Science, Cornell University, Ithaca, NY, Spring 1988.

[20] G. Salton and M. Smith, On the Application of Syntactic Methodologies in Automatic Text Analysis, *Proceedings of 12th International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, NY, June 1989, 137-150.

.I1
.W

Text Compression.

The usefulness and efficiency of text-processing systems can often be improved greatly by converting normal natural-language text representations into a new form better adapted to computer manipulation.

For example, storage space and processing time are saved in many applications by using short document abstracts, or summaries, instead of full document texts.

Alternatively, the texts can be stored and processed in encrypted form, rather than the usual format, to preserve the secrecy of the content.

One obvious factor usable in text transformations is the redundancy built into normal natural-language representation.

By eliminating redundancies--a method known as text compression, it is often possible to reduce text sizes considerably without any loss of text content.

Compression was especially attractive in earlier years, when computers of restricted size and capability were used to manipulate text.

Today large disk arrays are usually available, but using short texts and small dictionary sizes saves processing time in addition to storage space and still remains attractive.

.I 2
.W

Statistical Language Characteristics.

Frequency Considerations.

A study of the composition of text samples shows the uneveness with which linguistic text elements occur.

Thus for languages of Western Europe, the 15 alphabetic characters occurring most frequently in running text account for 85 to 95 percent of all letter occurrences. [1]

In English, the 10 most frequent alphabetic characters (**e, t, a, i, s, o, n, h, r,** and **d**) together account for over 70 percent of letter occurrences.

Letter combinations and complete words also occur unevenly. Thus many two-letter combinations (digrams) do not occur at all in English (e.g., **qz**), while others, such as **th** are quite frequent.

The 10 most frequent words in English, taken from a sample of over 1 million words of text, are shown in Table 5.1. [2]

It can be seen that the two words **the** and **of** account for over 10 percent of the word occurrences of English text, while the first six words (**the, or, and, to, a, in**) account for over 20 percent of all word occurrences in English.

The 50 most frequent words cover more than half of all word occurrences in ordinary text.

Two Sample Local Documents (Documents 1 and 2)

Fig. 1

Document 1

Syntactic phrases
text compression : R:1:A:1 S-1/1/1/1/
computer manipulation : R:1:A:1 P-0/1/1/2/
converting natural-language text representations : R:3:A:1 P-0/1/1/2/
efficiency : R:8:A:2 P-0/1/1/2/
systems : R:8:A:1 P-0/1/1/2/
usefulness : R:8:A:2 P-0/1/1/2/
document texts : R:1:A:1 P-1/1/2/3/
storage space : R:1:A:1 P-1/1/2/3/
summaries : R:8:A:1 P-1/1/2/3/
using document abstracts : R:3:A:1 P-1/1/2/3/
content secrecy : R:4:A:2 P-1/1/3/4/
encrypted : R:8:P:1:A:1 P-1/1/3/4/
format : R:8:A:2 P-1/1/3/4/
texts : R:8:A:2 P-1/1/3/4/
natural-language representation : R:1:P:1:A:1 P-1/1/1/5/
redundancy : R:8:A:2 P-1/1/1/5/
text transformations : R:1:P:1:A:1 P-1/1/1/5/
redundancies : R:8:A:1 P-1/1/2/6/
text compression : R:1:A:1 P-1/1/2/6/
text content loss : R:4:A:1 P-1/1/2/6/
text sizes : R:1:A:1 P-1/1/2/6/
compression : R:8:A:1 P-0/1/3/7/
computers : R:8:A:1 P-0/1/3/7/
restricted capability : R:2:A:1 P-0/1/3/7/
restricted size : R:2:A:1 P-0/1/3/7/
text : R:8:A:1 P-0/1/3/7/
addition : R:8:P:1:A:1 P-1/1/4/8/
saves : R:8:A:1 P-1/1/4/8/
storage space : R:1:A:1 P-1/1/4/8/
today disk arrays : R:1:A:1 P-1/1/4/8/
using small dictionary : R:3:A:1 P-1/1/4/8/
using texts dictionary : R:3:A:1 P-1/1/4/8/

Document 2

Syntactic phrases
statistical language characteristics : R:3:A:1 S-1/2/1/9/
frequency considerations : R:1:A:1 S-1/2/1/10/
composition : R:8:A:2 P-1/2/1/11/
linguistic text elements : R:3:A:1 P-1/2/1/11/
study : R:8:A:3 P-1/2/1/11/
text samples : R:1:A:1 P-1/2/1/11/
uneveness : R:8:A:2 P-1/2/1/11/
Western Europe languages : R:5:A:1 P-1/2/2/12/
alphabetic characters : R:2:A:2 P-1/2/2/12/
occurrences : R:8:A:1 P-1/2/2/12/
text account : R:1:A:1 P-1/2/2/12/
English : R:5:P:1:A:1 P-0/2/3/13/
account : R:8:A:1 P-0/2/3/13/
alphabetic characters : R:2:A:1 P-0/2/3/13/
letter occurrences : R:1:A:1 P-0/2/3/13/
letter combinations : R:1:A:1 P-1/2/1/14/
words : R:8:A:1 P-1/2/1/14/
English : R:5:A:1 P-0/2/2/15/
digrams : R:8:A:1 P-0/2/2/15/
two-letter combinations : R:1:A:1 P-0/2/2/15/
English : R:5:P:1:A:1 P-1/2/3/16/
Table : R:5:P:1:A:1 P-1/2/3/16/
frequent words : R:2:A:2 P-1/2/3/16/
sample : R:8:A:3 P-1/2/3/16/
text : R:8:A:1 P-1/2/3/16/
words : R:8:A:1 P-1/2/3/16/
English : R:5:P:1:A:1 P-0/2/4/17/
English text : R:5:A:1 P-0/2/4/17/
account : R:8:A:1 P-0/2/4/17/
occurrences : R:8:A:1 P-0/2/4/17/
word occurrences : R:1:A:2 P-0/2/4/17/
words : R:8:A:2 P-0/2/4/17/
frequent words : R:2:A:2 P-1/2/5/18/
ordinary text : R:2:P:1:A:1 P-1/2/5/18/
word occurrence half : R:4:A:1 P-1/2/5/18/

Unmodified Syntactic Indexing Units

Fig. 2

R marker: type of syntactic tag sequence
A marker: type of article preceding phrase
P marker: type of grammar used for parsing
Final digit sequence: location of phrase in text

Document 1				Document 2			
1	437	0.36273	text	2	1543	0.39658	occur
1	7652	0.24997	compress	2	7428	0.37442	engl
1	8796	0.24397	attract	2	8277	0.31450	word
1	879	0.22654	save	2	1504	0.30506	percent
1	3930	0.22654	redund	2	1999	0.28956	account
1	3259	0.22539	size	2	7454	0.28541	frequ
1	7612	0.17827	short	2	3145	0.26247	let
1	4611	0.16270	natur	2	1070	0.19213	unev
1	6264	0.16135	stor	2	437	0.15581	text
1	4855	0.15250	spac	2	381	0.13740	alphabet

a) Ten Best Terms in Book Context

	Document 1 Phrases		Document 2 Phrases
	text compress	(x)	text account
	stor spac		let occur
	text compress	(x)	frequ alphabet
	text size		let occur
	short text		occur unev
(x)	size save		frequ word
	stor spac		word occur
			engl text
			word occur
			frequ word
			word occur

b) Phrases Formed from Best Terms in Book Content

Document 1 Frequency Restricted Phrases		Document 2 Frequency Restricted Phrases	
stor space	(2)	frequ word	(2)
text compress	(2)	let occur	(2)
		word occur	(2)

c) Frequency Restricted Phrases

Term Value Computation and Phrase Formation

Table 1

	Chapter 5 (compression)	Chapter 6 (encryption)
Total statistical phrase occurrences	534	334
Total phrase occurrences with poor rating	95	50
Total phrase occurrences with acceptable rating	439	284
Proportion of acceptable phrase occurrences	0.822	0.850
Total statistical phrase occurrences with frequency exceeding 1 (f>1)	359	176
Phrase occurrences with poor rating (f>1)	50	17
Acceptable phrase occurrences (f>1)	309	159
Proportion of acceptable phrase occurrences (f>1)	0.849	0.903

Statistical Phrase Precision Computation

Table 2

	Chapter 5 (compression)	Chapter 6 (encryption)
Total number of automatically chosen statistical phrases	534	334
Total number of manually chosen relevant phrases	126	135
Average number of manually assigned relevant phrases per document	2.9	2.7
Statistical phrase-recall (proportion of relevant phrases occurring in statistical phrase set)	0.546	0.397
Total number of automatically chosen statistical phrases with frequency exceeding 1	359	176
Phrase-recall for frequency restricted statistical phrases	0.401	0.200

Statistical Phrase-Recall Computation

Table 3

	Chapter 5 (compression)	Chapter 6 (encryption)
Total number of unedited syntactic phrase occurrences	698	756
Total number of unedited syntactic phrases rated acceptable	422	556
Proportion of acceptable syntactic phrase occurrences	0.605	0.735
Total number of unedited syntactic single term occurrences	518	774
Total number of unedited syntactic single term occurrences rated acceptable	117	256
Proportion of acceptable syntactic single terms	0.226	0.331

Phrase Precision Computations for Unedited Syntactic Output

Table 4

Document		Phrases with R1-P1 and R1-S1 Constructions (all nouns or title phrases)	Phrases with R2-P1 Construction (adjective-noun combinations)
1		document texts	-
		natural-language representation	
		storage space (2)	
		text compression (2)	
		text sizes	
		text transformations	
	(x)	today disk arrays	
2		frequency considerations	alphabetic characters
		letter combinations	frequent words
		statistical language characteristics	ordinary text
	(x)	text account	
		text samples	

Phrases with Specified Syntactic Structure
(R1-P1, R1-S1, and R2-P1 Constructions)

Table 5

	Chapter 5 (compression)	Chapter 6 (encyrption)
Total number of syntactic phrases with R1-P1 or R1-S1 markers (all noun syntactic tags)	165	137
Proportion of syntactic phrases with R1-P1 or R1-S1 markers rated acceptable	0.921	0.920
Total number of syntactic phrases with R2-P1 markers (noun-adjective tags)	132	188
Proportion of syntactic phrases with R2-P1 markers rated acceptable	0.818	0.862
Total number of syntactic phrases with R1-P1 or R2-P1 markers	297	325
Proportion of R1-P1, R2-P1 phrase rated acceptable	0.875	0.886

Phrase Precision -- Proportion of Acceptable Syntactic Phrases with Syntactic Tag Restrictions

Table 6

Document		Phrases with all Components in Top-Ten Terms	Phrases with one Component in Top-Ten (Two Contexts)
1		document text storage space (2) text compression text size	natural-language representation restricted size text transformations
2		frequency considerations	letter combinations
	(x)	frequent words (2)	
		letter occurrences	
		statistical-language characteristics	
	(x)	text account	
		word occurrences	

Syntactic Phrases with Term Value Restrictions

Table 7

	Chapter 5 (compression)	Chapter 6 (encryption)
Total number of phrase occurrences with all components in top-ten	221	174
Proportion of phrase occurrences with acceptable rating	0.805	0.891
Total number of phrase occurrences with one component in top-ten	176	188
Proportion of such phrases rated acceptable	0.528	0.762
Total number of phrase occurrences with at least one component in top ten	397	362
Proportion of such phrases with acceptable rating	0.683	0.823

Phrase Precision Computation for Syntactic Phrases with Term Weight Restrictions

Table 8

	Chapter 5 (compression)	Chapter 6 (encyrption)
Total number of syntactic phrases with R1-P1 markers	165	137
Syntactic phrase-recall: Proportion of manually chosen relevant phrases among syntactic R1-P1 phrases	0.402	0.274
Total number of syntactic phrases with R1-P1 or R2-P1 markers	297	325
Syntactic phrase-recall: Proportion of manually chosen relevant phrases among syntactic R1-P1 and R2-P1 phrases	0.497	0.437
Total number of syntactic phrases with all components in top-ten	221	174
Syntactic phrase-recall: Proportion of manually chosen relevant phrases among syntactic phrases with highly weighted components	0.467	0.326
Total number of syntactic phrases with at least one component in top-ten	397	362
Syntactic phrase-recall: Proportion of manually chosen relevant phrases among syntactic phrases with at least one highly-weighted component	0.626	0.537

Phrase-Recall Computations for Syntactic Phrases

Table 9

		Chapter 5 (compression)	Chapter 6 (encryption)
a)	Statistical phrase occurrences		
	Number of acceptable statistical phrase occurrences	433	284
	Phrase-precision (proportion of acceptable statistical phrase occurrences)	0.822	0.850
	Number of acceptable statistical phrase occurrences (f>1)	309	159
	Phrase precision (proportion of acceptable statistical phrase occurrences (f>1))	0.849	0.903
b)	Syntactic phrase occurrences		
	Number of acceptable syntactic phrase occurrences with R1-P1 tags	152	126
	Phrase precision (proportion of acceptable syntactic phrase occurrences with R1-P1 or R2-P1 tags)	0.921	0.920
	Number of acceptable syntactic phrase occurrences with R1-P1 or R2-P1 tags	260	288
	Phrase precision (proportion of acceptable phrase occurrences with R1-P1 or R2-P1 tags)	0.875	0.886
c)	Overlap in acceptable statistical phrases (f>1) with acceptable R1-P1 + R2-P1 syntactic phrases	63/506 =0.125	51/396 =0.129

Phrase Precision Comparison for Syntactic and Statistical Phrases

Table 10