An Analysis of Statistical and Syntactic Phrases

Mandar Mitra*	Chris Buckley	Amit Singhal	Claire Cardie
Cornell University	Sabir Research Inc.	AT&T Research	Cornell University

Abstract

As the amount of textual information available through the World Wide Web grows, there is a growing need for high-precision IR systems that enable a user to find useful information from the masses of available textual data. Phrases have traditionally been regarded as precision-enhancing devices and have proved useful as content-identifiers in representing documents. In this study, we compare the usefulness of phrases recognized using linguistic methods and those recognized by statistical techniques. We focus in particular on high-precision retrieval. We discover that once a good basic ranking scheme is being used, the use of phrases does not have a major effect on precision at high ranks. Phrases are more useful at lower ranks where the connection between documents and relevance is more tenuous. Also, we find that the syntactic and statistical methods for recognizing phrases yield comparable performance.

1 Introduction

The amount of textual information available through the World Wide Web has increased dramatically in recent years. Users need effective search mechanisms in order to find useful information from the enormous quantities of available text data. Very often, Web users are precision-oriented — they prefer a small set of documents containing a good proportion of useful documents to a large set of documents that contains a lot of useful information, but a fair amount of irrelevant information as well. Thus, there has been a growing interest in high-precision IR systems in recent times.

One approach that has traditionally been regarded as a tool for increasing precision is the use of phrases for indexing and retrieval of documents[14, 16]. Consider the phrase "private investigator" taken from query 176 of the TREC collection [10].

> Topic: Real-life private investigators
>
> Description: Document must refer to the hiring of a private investigator in real-life or describe the work of a private investigator.
>
> Narrative: Private detectives can be self-employed or work for a private agency. Document can describe the work of private detectives in general without citing a specific person.

A document discussing private investigators would contain the single terms "private" and "investigator" and would be retrieved because of these matches even if phrases were not being used. Thus, the use of the phrase "private investigator" does not result in any new document being retrieved, i.e. total recall is not affected. Documents containing both the words "private" and "investigator" but in unrelated contexts would also have the same matching single terms, but would not contain the phrase. Such documents are likely to be non-relevant and the phrase match would promote the relevant documents above such non-relevant articles. This would enhance precision at the top ranks.

Phrases have been found to be useful indexing units by most of the leading groups participating at the NIST and DARPA sponsored Text REtrieval Conferences for performance evaluations of IR systems [10, 5, 4, 1, 11, 22, 21, 20, 7, 8]. In this study, we re-examine the usefulness of phrases, particularly within the context of high-precision retrieval.

Statistical and Syntactic Phrases. We consider two classes of phrases:

*This study was supported in part by the National Science Foundation under grant IRI-9624639.

1. Statistical phrases: any pair (or triple, quadruple, etc.) of non-function words that occur contiguously often enough in a corpus constitute a phrase. Thus, the words "United" and "States" may occur contiguously a large number of times in a corpus, and would constitute the phrase "United States".

2. Syntactic phrases: any set of words that satisfy certain syntactic relations or constitute specified syntactic structures make up a phrase. Thus, if we specify that an adjective followed by a noun constitutes a phrase, "economic impact" would be a phrase.

Syntactic phrases capture actual linguistic relations between words rather than the simple juxtaposition of words, and are expected to be semantically more meaningful. Our intention is to compare the usefulness of statistical and syntactic phrases, focusing our attention on noun phrases only.

Related Work. The relative merits of statistical and syntactic phrases were extensively investigated by Fagan [9]. This study used the very small CACM and CISI collections as document databases. A similar question was examined by Grefenstette et al. in [11], but the database used (the Wall Street Journal section of TREC disk 2) was much larger and more realistic. Similar comparisons are reported by Strzalkowski [20] also. Most of these studies use the standard Cornell Smart *lnc.ltc* [15] run as a baseline.

We continue to investigate the usefulness of phrases using a realistic database, but we start with a significantly improved baseline. A new term-weighting scheme developed by our group at Cornell University in [19, 18] outperforms the traditional *lnc.ltc* method by about 20% to as much as 43% for short queries [18, 5]. Several IR techniques that appear to work well with an inferior weighting scheme — massive expansion, and the combination of local and global similarities, for example — are no longer as useful once the basic term-weighting method is improved [5, 1]. In fact, a similar observation can be made about the use of phrases as well: in the past five years of TREC, overall retrieval effectiveness has more than doubled, but the added effectiveness due to statistical phrases has gone down from 7% to less than 1% [4]. The aim of this study is to examine the following questions in greater detail:

- Given a good basic document ranking scheme, what additional improvements can be obtained by using phrases in indexing and retrieval?
- Is there a significant difference in the benefits obtained from using syntactic vs. statistical phrases?

As explained above, we are specially interested in investigating these issues within the context of high-precision retrieval.

The rest of the paper is organized as follows: Section 2 presents the precise definitions of statistical and syntactic phrases used in this study; Section 3 describes the experimental methodology used and the results of our experiments; Section 4 discusses our findings; Section 5 concludes the paper.

2 Phrase Identification Methods

We use the Smart information retrieval system [13] in all our experiments. Smart is based on the vector space model of IR, which represents documents (and queries) by vectors of the form:

$$D_i = (w_{i1}, w_{i2}, \ldots, w_{it})$$

In the above expression, D_i represents a document (or query) text and w_{ik} is the weight of term T_k in document D_i. We also use the notation $w_d(T)$ to represent the weight of term T in document d. A weight of zero is used for terms that are absent from a particular document, and positive weights characterize terms contained in a document. The assumption is that t terms in all are available for the representation of the information.

The weight assigned to a term usually takes into account the following factors:

- *term frequency* or tf: the number of occurrences of the term within the document,
- *inverse document frequency* or idf: an inverse measure of the number of documents in the collection in which the term occurs, and

- *normalization factor*: the length of the document.

A triple of letters is used to denote the actual formula used to assign weights to terms in a particular experiment [15]. The first letter indicates how the tf factor is handled; the second letter indicates how idf information is incorporated; the third letter gives the normalization factor. A retrieval run is designated by a pair of such triples, the first indicating the term-weighting formula used for document term weights and the second indicating the formula used for query term weights.

Linguistic processing is done using the Circus system [6, 12]. The IR and NLP subsystems interact as follows: Smart provides Circus with the text of a document or query; Circus identifies the noun phrases contained in that text and returns a list of such phrases to Smart; Smart converts the list of phrases into its vector representation and uses this in all subsequent processing.

Statistical Phrases. The definition of a statistical phrase is very simple. All pairs of non-function words that occur contiguously in at least 25 documents in disk 1 of the TREC collection are regarded as phrases. The individual words are stemmed and the pair is ordered lexicographically (thus, "United States" becomes "stat unit") to obtain the final form of the phrase that is used to index a document.

Syntactic Phrases. To identify the syntactic phrases present in a document, we first tag every word in the document with its part of speech (POS) using the Brill tagger [3, 2]. Certain tag patterns are then recognized as noun phrases (NPs). The following grammar specifies the POS tag patterns that are recognized as phrases[1].

NP → NN
→ NNS
→ NNP
→ NNPS
→ NN NP
→ NNS NP
→ NNP NP
→ NNPS NP
→ PRP$ NP
→ DT NP
→ JJ NP
→ JJR NP
→ JJS NP
→ CD NP
→ NP POSNP
→ VBG NP
POSNP → POS NP

The NLP subsystem returns a list of the maximal noun phrases found in a document. A maximal noun phrase is one that is not a constituent of a longer noun phrase according to the above rules. For example, consider the sentence "Information about the acid rain problem is irrelevant". The phrase "*the acid rain problem*" is identified as an NP by the above rules, but no phrase containing it is recognized as an NP. This NP is therefore a maximal noun phrase. Similarly, the NP "*acid rain*" is not a maximal noun phrase since it is contained within "*the acid rain problem*".

Each phrase is stripped of all stopwords and the non-stop words are stemmed to yield the final form of the phrase that is used to index the document[2]. Further, for phrases consisting of three or more single words, all pairs of single words are added as additional phrases. For example, the phrase "the National Rifle

[1] The symbols used in the grammar correspond to the following parts of speech: CD – cardinal number, DT – determiner, JJ – adjective, JJR – comparative adjective, JJS – superlative adjective, NN – singular or mass noun, NNP – singular proper noun, NNPS – plural proper noun, NNS – plural noun, POS – possessive ending, PRP$ – possesive pronoun, VBG – verb, gerund or present participle.

[2] We also tried indexing documents by the original phrases without stemming and stopword removal but did not find that useful.

Association" is stripped down to "nation rifl assoc" and yields the phrases "nation rifl assoc", "nation rifl", "nation assoc", and "rifl assoc".

Terms in documents are weighted by *Lnu* weights and query terms are weighted using the *ltu* scheme [18]. For single terms and statistical phrases, these weighting methods have their usual significance under the Smart triple notation. The idf factor (given by $\log(N/df)$, where N is the number of documents in the collection and df is the number of documents containing the given term) has to be calculated differently for syntactic phrases, however. Since the process of identifying the syntactic phrases present in a document is computationally expensive, not all documents in the collection are indexed by syntactic phrases (see Section 3). Consequently, we do not know the true document frequencies (or dfs) for these phrases, and some approximation must be used. We try the following methods for estimating the idf of syntactic phrases based on the idfs of constituent single words:

1. Maximum: The idf of a phrase is the maximum of the idfs of constituent single words.
2. Minimum: This is defined analogously.
3. Arithmetic mean: The idf of a phrase is the arithmetic mean of the idfs of the constituent single words.
4. Geometric mean: This is defined analogously.
5. Intersection (represented by 'T' in the triple notation): The document frequency of a phrase is the number of documents in which all of its constituent words occur. Note that this approximate formulation does not take into account whether the constituent words actually occur in a phrasal relation.
6. Probabilistic (represented by 'P' in the triple notation): For a two-word phrase $\mathcal{P} = A\ B$, the idf of $\mathcal{P}$ is given by:

$$\log(N/N_{AB}) \times -\log(\frac{N_{AB}}{N} - \frac{N_A \times N_B}{N_{AB}^2})$$

   if $\mathcal{C} = \frac{N_{AB}}{N} - \frac{N_A \times N_B}{N_{AB}^2}$ is positive, and is set to zero otherwise. Here N is the number of documents in the collection, and N_A, N_B, N_{AB} are respectively the number of documents containing the term A, B, and both terms.

   This formulation can be motivated as follows. Consider a phrase like "Los Angeles" or "United States". A query-document match on the word "Angeles" indicates that the query-document pair is very likely to match on the phrase "Los Angeles" as well. In this case, the phrase match does not provide much additional information about the document. In contrast, consider a phrase like "dangerous toy" — the word "dangerous" occurs in several contexts unrelated to "toy" and vice versa. A query-document match on "dangerous toy" provides us with more information than a match on either of the single words. We would therefore like to increase the importance of phrases like "dangerous toy" compared to phrases like "Los Angeles". This is accomplished by incorporating an inverse function of $\mathcal{C}$ — $\mathcal{C}$ is high for "Los Angeles" and low for "dangerous toy" — in the term-weights. The method seems to work well overall as shown below, and needs further investigation.

3 Experiments and Results

The document database used in our experiments consists of the Wall Street Journal, AP Newswire, and Ziff-Davis sections of TREC disk 2, a total of 211359 documents. The query collection comprises TREC queries 151 – 200. The number of relevant documents in the collection for this query set is 4273.

One straightforward way to evaluate the usefulness of phrases, and to compare the performance of syntactic and statistical phrases would be to retrieve a certain number of documents (1000, say) for each of the 50 queries, and compare the results obtained using phrases to the results when phrases are not used. This would require that all documents in the collection be indexed for syntactic phrases. As mentioned in the previous section, this is an expensive process. In order to avoid the time-consuming step of identifying the syntactic phrases for every document in the collection, we adopt the following approach. Using single term matches only, 100 documents are retrieved for each of the 50 queries. These 100 documents are then indexed by syntactic phrases, and phrase matches are used to rerank these 100 documents. The final similarity scores for these documents is given by the following expression:

$$Sim_{final} = Sim_{old} \; + \; (\textbf{phrase-factor} \times \sum_{t \in matching\ phrases} w_{doc}(t) * w_{query}(t))$$

Sim_{old} is the similarity based on matching single terms and is used to rank the initially retrieved set of 100 documents. The second summand gives the similarity based on matching phrases. **phrase-factor** is used to vary the importance given to phrase matches. Typically, a value of 0.5 is used for statistical phrases[3].

The initial retrieval run based on single term matches is treated as the baseline, and the results of reranking the top-ranked documents using phrase matches are compared against this run. As explained in the Introduction, phrases are generally regarded as precision-enhancing devices. Thus, this approach of studying rank changes caused by phrase matches among the top-ranked documents appears to be a reasonable one.

Our initial experiments were designed to study the behavior of syntactic phrases in order to answer questions such as

- How useful are three-word (or longer) phrases?
- What weighting scheme works well with syntactic phrases?
- Should phrases be stemmed and have stopwords removed?

Once an "optimal" way to use syntactic phrases is developed, we compare the performance of these phrases with that of statistical phrases.

Three-word phrases. A total of 527 syntactic phrases are recognized for all the 50 queries. Of these, 29 consist of at least three words. All others are two-word phrases. Since we limit ourselves to considering two-word statistical phrases only, we need to study the effect of a similar limitation in the case of syntactic phrases. Accordingly, we first compared a run that used all syntactic phrases with one that used only two-word phrases and two-word sub-phrases of longer phrases. The average precision obtained using various idf weighting schemes are shown in Table 1 (recall that we use $Lnu.l(\cdot)u$ weights). For each scheme, a phrase-factor of 0.5 is used. This turns out to be a good value to use.

Idf factor	Maximum	Minimum	Arithmetic Mean	Geometric Mean
All phrases	0.2939	0.2938	0.2941	0.2944
Two-word phrases	0.2943	0.2939	0.2948	0.2948

Table 1: Comparison of all phrases and two-word phrases.

The differences are fairly small across various weighting methods. In fact, performance improves very slightly when we limit ourselves to using two-word phrases only. This can be explained as follows: a match on all two-word sub-phrases of a long phrase is almost always equivalent to a match on the entire phrase and in such a situation, the match on the entire phrase does not provide additional information. Further, when a matching long phrase (together with all the matching sub-phrases) contributes to the query-document similarity, the importance of this single match maybe over-emphasized and this hurts performance [4]. If the phrase-factor were reduced (to 0.25, say), the problem of over-emphasizing a phrase match would no longer be as serious, since phrase matches in general would be given lower importance and we would expect the difference between using all phrases and using two-word phrases only to diminish. This is what we actually observed in the course of our experiments, and this provides further support to our belief. In the following, therefore, we consider two-word phrases only.

[3] This value has been used for a long time at Cornell. Experiments by Grefenstette et al. [11] also conclude that operationally, this is a good value. Our experiments with the current test database confirms this.

[4] In fact, the same rationale motivated the use of two-word statistical phrases.

Phrase weighting. In addition to the idf weighting schemes shown in Table 1, we also tried the schemes "T" and "P" for two-word phrases (with a phrase-factor of 0.25 and 0.05 respectively[5]). The results for the various methods are shown in Table 2.

Idf factor	Maximum	Minimum	Arithmetic Mean	Geometric Mean	T	P
Avg. Precision	0.2943	0.2939	0.2948	0.2948	0.2949	0.2957

Table 2: Performance of various phrase weighting schemes.

The differences between various weighting methods are negligible. For our subsequent experiments, we use the *lTu* and *lPu* schemes for query term weights. The *lTu* scheme is a reasonable approximation to using true idf weights for phrases. The *lPu* scheme is very similar to the *lTu* scheme, and for most queries, it does marginally better, giving improved results for 30 out of the 50 queries.

Statistical and syntactic phrases. Table 3 compares the results of reranking the top documents using phrases to the base run. The last column shows the results if statistical phrases were used in the retrieval process itself rather than in reranking a set of documents retrieved through single term matches.

	Base run	Statistical phrases	Syntactic phrases		Retrieval with
			lTu	lPu	terms + stat. phr.
Avg. precision	0.2925	0.2929	0.2949	0.2957	0.3020
Precision at 20 docs.	0.5420	0.5480	0.5480	0.5510	0.5500

Table 3: Comparison of single terms, statistical, and syntactic phrases.

From the table it is immediately clear that changes observed are very small. Thus for example, reranking the top 100 documents using statistical phrases yields an improvement of 0.1% only. The improvement is marginally more (0.8% to 1.1%) for syntactic phrases, but this is not significant. Table 3 also shows that the average number of relevant documents in the top 20 remains almost unaffected by the use of phrases: on an average, there are 10.84 relevant documents in the top 20 in the base results and 11.02 relevant documents when syntactic phrases are used.

	Syntactic (lTu) better	Statistical better	Syntactic (lPu) better	Statistical better
Avg. Precision	12	12	14	10
Precision at 20 docs.	6	5	6	5

Table 4: Number of queries for which statistical and syntactic phrases yield significantly different performance.

Table 4 shows the number of queries for which the performance of syntactic and statistical phrases is noticeably different. In terms of average precision, the performance of statistical and syntactic phrases differs significantly for only about half the queries — for 12 queries, syntactic phrases (using *lTu* weights) outperform statistical phrases by at least 5%, for 12 queries, they are worse by at least 5%. Looking at the comparative performance of the two methods in terms of precision at a 20 document cutoff, we find that the difference in the number of relevant documents is at least two for only 11 queries: for 5 of them, statistical phrases do better, and for the remaining 6, syntactic phrases give better precision.

[5] The additional multiplicative factor used in the "P" scheme scales up the weights of all phrases. Hence a lower phrase-factor is needed with this scheme.

Comparing the set of statistical and syntactic phrases used to index the 50 queries, we find that there are 160 phrases which are identified by the statistical method alone, 214 phrases identified by the syntactic method alone, and 269 phrases are picked up by both methods. Since over half the phrases are shared, it is not surprising that many queries are not strongly affected by the choice of phrase strategy. But for the remaining queries it is clear that neither syntactic nor statistical phrases have an advantage over the other.

The differences between syntactic and statistical phrases are further examined in Table 5. We reranked the top 100 documents based on phrase matches only (this is equivalent to using a phrase-factor of infinity). Phrases by themselves are non-optimal representatives of a query, since normally only a part of the subject matter of a query is covered by phrases. Consequently, the results shown below are considerably worse than the baseline, but they are useful since they show the effects of phrases in isolation. The benefits of using linguistic information to recognize phrases are clear from the table: the *lPu* weighted syntactic phrases are 10% better than the statistical phrases in terms of precision at 20 documents. We should consider this improvement in the context of Table 3 however, which indicates that these benefits may be of much less importance within a good indexing and retrieval system.

	Base run	Statistical phrases	Syntactic phrases	
			lTu	lPu
Avg. precision	0.2925	0.2401	0.2502	0.2521
Precision at 20 docs.	0.5420	0.4590	0.4860	0.5030

Table 5: Results of reranking using phrase matches only.

4 Discussion

The smallness of the changes produced by the variations we tried, and the closeness of the number of queries that are improved or hurt by either technique leads us to conjecture that the top ranks is not where phrases are most useful. The following observations support this conjecture.

Phrase matches at high ranks. As explained in the Introduction, the generally held belief that phrases improve precision is based on the hypothesis that a phrase match would promote relevant documents above non-relevant documents in which two query words occur, but not in the intended sense or semantic relation. A closer look at the top-ranked non-relevant documents shows that such non-relevant documents are not very common. What seems to be the reason for low precision in a large number of cases is that the query consists of several aspects, and some top-ranked documents deal with only one of these aspects and are therefore not relevant to the entire query. For example, query 184 deals with problems associated with pension plans/funds such as fraud, skimming, tapping or raiding. Several top-ranked documents discuss pension plans but do not deal with associated problems.

The problem in such cases can be termed as one of inadequate query coverage [4]. The use of phrases does not consistently address this problem: in a query with multiple aspects, a phrase normally does not capture the multiple aspects of the query, but deals with one particular aspect only. If this happens to be the main aspect, but is not represented well by single terms, then the use of phrases helps. On the other hand, if this aspect is already the dominant retrieval component due to the single terms (i.e., most top-ranked documents address this aspect), the use of phrases only over-emphasizes the aspect. For example, for query 176 on real-life private investigators, the phrase “private investigator” is vital, and the use of phrases improves precision for this query. Conversely, query 165 (“Tobacco company advertising and the young”) quite clearly has multiple aspects. Most of the top-ranked documents deal with the tobacco industry, but not necessarily with the effect of the industry’s advertising on youth. Phrases such as “tobacco industry” and “tobacco company” emphasize an already dominant component and causes precision to drop. Query 184, mentioned in the previous paragraph, is another example in this category.

These observations indicate that phrases cannot be depended on as consistent precision-enhancing devices and suggest that in order to improve precision at high ranks, we need to use methods that directly address the coverage problem. This also introduces another wrinkle into the comparison of statistical and syntactic phrases. It shows the important consideration is not whether syntactic phrases are better at retrieval than statistical phrases (Table 5 shows they are), but whether they give different information than the single terms do.

Phrase matches at low ranks. If, as the above indicates, phrases do not help in a high-precision, low-recall scenario, we are led to ask whether phrases help in improving the relative ranking for low-ranked documents. Running a full retrieval run where both statistical phrases and single terms are used to index documents, we obtain an improvement of 3.2% over the baseline (see the last column in Table 3). This is much greater compared to the improvements obtained using phrases only to rerank the top-ranked documents. This indicates that the reranking effect that phrases have at relatively poor ranks could well be more beneficial. These rerankings would result in relevant documents ranked below 100 by single term matches being pulled into the top 100 when phrases are used.

To corroborate this surmise, we reranked the top 500 documents instead of the top 100 only. If phrases are more useful at low ranks than at the top ranks, the improvements from reranking using phrase matches should be more noticeable when 500 top-ranked documents are reranked rather than 100. Table 6 shows that this is indeed the case. The precision at a 20 document cutoff remains unchanged, but the improvement in average precision becomes more noticeable. Similar results are observed when phrases are used during retrieval. This run retrieves an additional 100 relevant documents (3061 compared to 2962 relevant documents retrieved by the base run using single terms only), but the precision at low recall levels does not change appreciably. This indicates that these relevant documents are being brought into the retrieved set at low ranks via the rerankings produced by the phrase matches around the lower edge of the retrieved pool.

The differences between statistical and syntactic phrases is still very small, however — syntactic phrases perform somewhat better in the high-precision region, and somewhat worse in the high-recall region.

	Base run	Statistical phrases	Syntactic phrases		Retrieval with terms + phrases
			lTu	lPu	
Avg. precision	0.3616	0.3710	0.3678	0.3697	0.3758
Precision at 20 docs.	0.5420	0.5500	0.5480	0.5510	0.5500

Table 6: Results of reranking 500 top-ranked documents.

We analyze the results for syntactic phrases in greater detail in Table 7, which shows the interpolated precision at various recall levels for both lTu and lPu weighted syntactic phrases. The trends visible from this table provide further support to our hypothesis that phrases help improve rankings at low ranks rather than at high ranks. We observe that as the importance given to phrase matches is increased, the precision at top ranks consistently deteriorates, but precision at lower ranks (or higher recall levels) consistently improves. In fact, the improvements at the high recall levels sometimes outweigh the performance losses at the top ranks when 500 documents are being reranked, and better overall performance is achieved by using a higher value of the phrase-factor compared to the values that are best for reranking the top 100 documents (0.25 for lTu-weighting and 0.05 for lPu-weighting).

Very similar patterns can be seen in the case of statistical phrases as well, when the phrase-factor is increased to 0.75 and 1.0. The same trends are also observable when 100 documents are reranked — performance at high recall levels improves as the phrase-factor is increased.

One possible explanation of this would be that at high ranks, the number of single term matches is typically high and the overlap between the subject matter of the query and document is substantial. Emphasizing a phrase match in this setting would over-emphasize a particular match, and cause retrieval to be dominated by this match. The resulting top-ranked documents could give a one-sided view of the information required by the query. At lower ranks, the number of matches is much lower and given the small number of matches, the importance of a good match (a phrase match for example) compared to an inferior match in distinguishing a good document from a bad one increases significantly. Accurately matching a document and query on a single aspect using a phrase match becomes more important here, since all of the matches at lower ranks are more tenuous.

Recall level	Phrase-Factor						
	lTu			lPu			
	0.25	0.50	1.0	0.05	0.10	0.15	0.25
at 0.00	0.8528	0.8416	0.8512	0.8516	0.8293	0.8384	0.8338
at 0.10	0.6986	0.6970	0.6817	0.7003	0.7004	0.6941	0.6799
at 0.30	0.5378	0.5328	0.5143	0.5419	0.5360	0.5315	0.5210
at 0.70	0.2090	0.2166	0.2256	0.2125	0.2171	0.2235	0.2255
at 1.00	0.0076	0.0095	0.0111	0.0082	0.0101	0.0109	0.0114
Avg. precision	0.3678	0.3694	0.3665	0.3697	0.3716	0.3715	0.3654

Table 7: Effect of phrase-factor at various recall levels.

This explanation is analogous to the one given by Singhal [17] to explain how using a strong inverse function of the document frequency ($(\log(N/df))^{1.5}$ for example) to compute the idf component of term-weights improves performance at high recall levels. Singhal broadly classifies query terms into two groups — rare (or high idf) terms which typically form the core of the topic, and less rare terms which play a supporting role by specifying the context or subtopic within the main topic that is of interest to the user. Top-ranked documents typically contain the core as well as some (or most) of the supporting terms. Using a strong idf function boosts the importance of the core term match and a non-relevant document that contains the core term but lacks the supporting terms may be promoted over a relevant document that contains both types of terms. In the lower ranks, however, the query-document match is weaker, and rare terms are a more reliable indicator of relevance, since the chances of relevance are low if the core topic is absent from a document, whereas a document may still be of interest if some of the supporting terms are missing. Using a strong idf function is useful here, since it gives more importance to a single core match compared to a number of matches of poorer quality.

5 Conclusion

Phrases have proved to be useful as indexing units in IR systems in the past. A phrase match between a document and a query is usually an accurate indication that the document deals with the aspect of the query described by the phrase. The ability to accurately detect an overlap between a document and a query on a single query aspect is important if retrieval performance is low. But if good performance is already achieved using single terms, adding phrases of any kind is likely to over-emphasize a particular aspect of a query, resulting in poorer performance. Thus, we observe that as overall retrieval effectiveness improves, the additional benefits obtained through the use of phrases seems to be diminishing. In the past five years of TREC, overall retrieval effectiveness for Smart has more than doubled, but the added effectiveness due to statistical phrases has gone down from 7% to less than 1%.

We examine this issue in this study, investigating the usefulness of phrases in general, and the comparative usefulness of statistical and syntactic phrases in particular, within a high-performance IR system. We are specially interested in the problem of high-precision retrieval.

We conclude the following from the results of our experiments:

- When phrase matches alone are used to rank documents, syntactic phrases perform better than statistical phrases, but this advantage disappears when single terms are used in indexing and retrieval.
- On average, the use of phrases does not significantly affect precision at the top ranks. Preliminary observations indicate that phrases are more useful in determining the relative ranks of low-ranked documents.
- Phrases are useful for some queries, but not others. The major issue seems to be one of query accuracy (increased by phrases) versus query coverage (with query balance possibly upset by phrases). The tradeoffs between these two factors need to be explored in the future.

6 Acknowledgments

The grammar used to recognize noun phrases was designed by Julia Komissarchik. Kevin Saunders was of great help in getting Circus to work on the text collections used in this study.

References

[1] J. Allan, L. Ballesteros, J. P. Callan, W. B. Croft, and Z. Lu. Recent Experiments with INQUERY. In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*. NIST Special Publication 500-236, October 1996.

[2] E. Brill. *A Corpus-Based Approach to Language Learning.* PhD thesis, University of Pennsylvania, Philadelphia, PA, 1993.

[3] E. Brill. Some Advances in Transformation-Based Part of Speech Tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 722–727, 1994.

[4] C. Buckley, A. Singhal, and M. Mitra. Using Query Zoning and Correlation within SMART: TREC5. In D. K. Harman, editor, *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*.

[5] C. Buckley, A. Singhal, M. Mitra, and (G. Salton). New Retrieval Approaches using SMART: TREC-4. In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*. NIST Special Publication 500-236, October 1996.

[6] C. Cardie and W. Lehnert. A Cognitively Plausible Approach to Understanding Complicated Syntax. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 117–124, 1991.

[7] D. A. Evans and R. G. Lefferts. CLARIT-TREC Experiments. *Information Processing and Management*, 31(3):385–395, 1995.

[8] D. A. Evans, N. Milic-Frayling, and R. G. Lefferts. CLARIT TREC-4 Experiments. In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*. NIST Special Publication 500-236, October 1996.

[9] J. L. Fagan. *Experiments in Automatic Phrase Indexing For Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods.* PhD thesis, Department of Computer Science, Cornell University, Ithaca, NY 14853, 1987. Available as CUCS Technical Report TR87-868.

[10] D. K. Harman. Overview of the Fourth Text REtrieval Conference (TREC-4) . In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*. NIST Special Publication 500-236, October 1996.

[11] D. A. Hull, G. Grefenstette, B. M. Schulze, E. Gaussier, H. Schutze, and J. O. Pedersen. Xerox TREC-5 Site Report: Routing, Filtering, NLP, and Spanish Tracks. In D. K. Harman, editor, *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*.

[12] W. Lehnert. Symbolic/Subsymbolic Sentence Analysis: Exploiting the Best of Two Worlds. In J. Barnden and J. Pollack, editors, *Advances in Connectionist and Neural Computation Theory*, pages 135–164. Ablex Publishers, Norwood, NJ, 1990.

[13] G. Salton, editor. *The SMART Retrieval System—Experiments in Automatic Document Processing.* Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

[14] G. Salton. *Automatic Text Processing—the Transformation, Analysis and Retrieval of Information by Computer.* Addison-Wesley Publishing Co., Reading, MA, 1989.

[15] G. Salton and C. Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[16] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw Hill Book Co., New York, 1983.

[17] A. Singhal. *Term Weighting Revisited.* PhD thesis, Department of Computer Science, Cornell University, Ithaca, NY 14853, 1996.

[18] A. Singhal, C. Buckley, and M. Mitra. Pivoted Document Length Normalization. In P. Schauble R. Wilkinson H. Frei, D. Harman, editor, *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

[19] A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document Length Normalization. *Information Processing and Management* (to appear). Also Technical Report TR95-1529, Department of Computer Science, Cornell University, Ithaca, NY 14853, July 1995.

[20] T. Strzalkowski, L. Guthrie, J. Karlgren, J. Leistensnider, F. Lin, J. Perez-Carballo, T. Straszheim, J. Wang, and J. Wilding. Natural Language Information Retrieval: TREC-5 Report. In D. K. Harman, editor, *Proceedings of the Fifth Text REtrieval Conference (TREC-5).*

[21] T. Strzalkowski and J. Perez-Carballo. Natural Language Information Retrieval: TREC-4 Report. In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4).* NIST Special Publication 500-236, October 1996.

[22] T. Strzalkowski, J. Perez-Carballo, and M. Marinescu. Natural Language Information Retrieval: TREC-3 Report. In D. K. Harman, editor, *Proceedings of the Third Text REtrieval Conference (TREC-3).* NIST Special Publication 500-225, April 1996.