1,700 of the three-word phrases are attested in the Lexile corpus. 14,045 of the 19,939 attested two-word phrases occur at least 5 times, 11,384 occur at least 10 times, and only 5,366 occur at least 50 times; in short, the strategy of cutting off the data at a threshold sacrifices a large percent of total recall. Thus one of the issues that needs to be addressed is the accuracy with which lexical association measures can be extended to deal with relatively sparse data, e.g., phrases that appear less than ten times in the source corpus.

A second question of interest is the effect of filtering for particular linguistic patterns. This is another method of prescreening the source data which can improve precision but damage recall. In the evaluation bigrams were classified as N-N and A-N sequences using a dictionary template, with the expected effect. For instance, if the WordNet two word phrase list is limited only to those which could be interpreted as noun-noun or adjective noun sequences, N>=5, the total set of WordNet terms that can be retrieved is reduced to 9,757..

4 Evaluation

Schone and Jurafsky's (2001) study examined the performance of various association metrics on a corpus of 6.7 million words with a cutoff of N=10. The resulting n-gram set had a maximum recall of 2,610 phrasal terms from the WordNet gold standard, and found the best figure of merit for any of the association metrics even with linguistic filterering to be 0.265. On the significantly larger Lexile corpus N must be set higher (around N=50) to make the results comparable. The statistics were also calculated for N=50, N=10 and N=5 in order to see what the effect of including more (relatively rare) n-grams would be on the overall performance for each statistic. Since many of the statistics are defined without interpolation only for bigrams, and the number of WordNet trigrams at N=50 is very small, the full set of scores were only calculated on the bigram data. For trigrams, in addition to rank ratio and frequency scores, extended pointwise mutual information and true mutual information scores were calculated using the formulas log $(P_{xyz}/P_xP_y\ P_z))$ and P_{xyz} log $(P_{xyz}/P_xP_y\ P_z))$. Also, since the standard lexical association metrics cannot be calculated across different n-gram types, results for bigrams and trigrams are presented separately for purposes of comparison.

The results are are shown in Tables 2-5. Two points should should be noted in particular. First, the rank ratio statistic outperformed the other association measures tested across the board. Its best performance, a score of 0.323 in the part of speech filtered condition with N=50, outdistanced

METRIC	POS Filtered	Unfiltered
RankRatio	0.323	0.196
Mutual Expectancy	0.144	0.069
TMI	0.209	0.096
PMI	0.287	0.166
Chi-sqr	0.285	0.152
T-Score	0.154	0.046
C-Values	0.065	0.048
Frequency	0.130	0.044

Table 2. Bigram Scores for Lexical Association Measures with N=50

METRIC	POS Filtered	Unfiltered
RankRatio	0.218	0.125
MutualExpectation	0.140	0.071
TMI	0.150	0.070
PMI	0.147	0.065
Chi-sqr	0.145	0.065
T-Score	0.112	0.048
C-Values	0.096	0.036
Frequency	0.093	0.034

Table 3. Bigram Scores for Lexical Association Measures with N=10

METRIC	POS Filtered	Unfiltered
RankRatio	0.188	0.110
Mutual Expectancy	0.141	0.073
TMI	0.131	0.063
PMI	0.108	0.047
Chi-sqr	0.107	0.047
T-Score	0.098	0.043
C-Values	0.084	0.031
Frequency	0.081	0.021

Table 4. Bigram Scores for Lexical Association Measures with N=5

METRIC	N=50	N=10	N=5
RankRatio	0.273	0.137	0.103
PMI	0.219	0.121	0.059
TMI	0.137	0.074	0.056
Frequency	0.089	0.047	0.035

Table 5. Trigram scores for Lexical Association Measures at N=50, 10 and 5 without linguistic filtering.