C E N T E R F O R R E S E A R C H I N L A N G U A G E

December 2003 **Vol. 15, No. 2**

The Newsletter of the Center for Research in Language, University of California, San Diego, La Jolla CA 92093-0526
Tel: (858) 534-2536 • E-mail: editor@crl.ucsd.edu • WWW: http://crl.ucsd.edu/newsletter

• • •

FEATURE ARTICLE

New Corpora, New Tests, and New Data for Frequency-Based Corpus Comparisons

Robert Liebscher

Department of Cognitive Science
University of California, San Diego

EDITOR'S NOTE

This newsletter is produced and distributed by the **CENTER FOR RESEARCH IN LANGUAGE**, a research center at the University of California, San Diego that unites the efforts of fields such as Cognitive Science, Linguistics, Psychology, Computer Science, Sociology, and Philosophy, all who share an interest in language. We feature papers related to language and cognition distributed via the World Wide Web) and welcome response from friends and colleagues at UCSD as well as other institutions. Please visit our web site at `http://crl.ucsd.edu`.

SUBSCRIPTION INFORMATION

If you know of others who would be interested in receiving the newsletter, you may add them to our email subscription list by sending an email to `majordomo@crl.ucsd.edu` with the line `"subscribe newsletter <email-address>"` in the body of the message (e.g., `subscribe newsletter jdoe@ucsd.edu`). Please forward correspondence to:

Ayşe Pınar Saygın and Jenny Staab, Editors
Center for Research in Language, 0526
9500 Gilman Drive, University of California, San Diego 92093-0526
Telephone: (858) 534-2536 • E-mail: editor@crl.ucsd.edu

Back issues of this newsletter are available on our website. Papers featured in recent issues include the following:

Rapid Word Learning by 15-Month-Olds under Tightly Controlled Conditions
Graham Schafer and Kim Plunkett
Experimental Psychology, Oxford University
Vol. 10, No. 5, March 1996

Learning and the Emergence of Coordinated Communication
Michael Oliphant and John Batali
Department of Cognitive Science, UCSD
Vol. 11, No. 1, February, 1997

Contexts That Pack a Punch: Lexical Class Priming of Picture Naming
Kara Federmeier and Elizabeth Bates
Department of Cognitive Science, UCSD
Vol. 11, No. 2, April, 1997

Lexicons in Contact: A Neural Network Model of Language Change
Lucy Hadden
Department of Cognitive Science, UCSD
Vol. 11, No. 3, January, 1998

On the Compatibility of CogLexicons in Contact: A Neural Network Model of Language Change
Mark Collier
Department of Philosophy, UCSD
Vol. 11, No. 4, June, 1998

Analyzing Semantic Processing Using Event-Related Brain Potentials
Jenny Shao, Northwestern University
Helen Neville, University of Oregon
Vol. 11, No. 5, December 1998

Blending and Your Bank Account: Conceptual Blending in ATM Design
Barbara E. Holder
Department of Cognitive Science, UCSD
Vol. 11, No. 6, April 1999

Could Sarah Read the Wall Street Journal?
Ezra Van Everbroeck
Department of Linguistics, UCSD
Vol. 11, No. 7, November 1999

Introducing the CRL International Picture-Naming Project (CRL-IPNP)
Elizabeth Bates, et al.
Vol. 12, No. 1, May 2000

Objective Visual Complexity as a Variable in Studies of Picture Naming
Anna Székely
Eotvos Lorand University, Budapest
Elizabeth Bates
University of California, San Diego
Vol. 12, No. 2, July 2000

The Brain's Language
Kara Federmeier and Marta Kutas
Department of Cognitive Science, UCSD
Vol. 12, No.3, November 2000

The Frequency of Major Sentence Types over Discourse Levels: A Corpus Analysis
Frederic Dick and Jeffrey Elman
Department of Cognitive Science, UCSD
Vol. 13, No.1, February 2001

A Study of Age-of-Acquisition (AoA) Ratings in Adults
Gowri K. Iyer, Cristina M. Saccuman, Elizabeth A. Bates, and Beverly B.Wulfeck
Language & Communicative Disorders, SDSU & UCSD and Center for Research in Language, UCSD
Vol. 13, No. 2, May 2001

Syntactic Processing in High- and Low-skill Comprehenders Working under Normal and Stressful Conditions
Frederic Dick, Department of Cognitive Science, UCSD
Morton Ann Gernsbacher, Department of Psychology, University of Wisconsin
Rachel R. Robertson, Department of Psychology, Emory University
Vol. 14, No. 1, February 2002

Teasing Apart Actions and Objects: A Picture Naming Study
Analia L. Arevalo
Language & Communicative Disorders, SDSU & UCSD
Vol. 14, No. 2, May 2002

The Effects of Linguistic Mediation on the Identification of Environmental Sounds
Frederic Dick , Joseph Bussiere and Ayşe Pınar Saygın
Department of Cognitive Science and Center for Research in Language, UCSD
Vol. 14, No. 3, August 2002

On the Role of the Anterior Superior Temporal Lobe in Language Processing: Hints from Functional Neuroimaging Studies
Jenny Staab
Language & Communicative Disorders, SDSU & UCSD
Vol. 14, No. 4, December 2002

A Phonetic Study of Voiced, Voiceless, and Alternating Stops in Turkish
Stephen M. Wilson
Neuroscience Interdepartmental Program, UCLA
Vol. 15, No. 1, April 2003

New Corpora, New Tests, and New Data for Frequency-Based Corpus Comparisons *

Robert Liebscher

Department of Cognitive Science
University of California, San Diego

Abstract

This study presents new data on frequency based corpus comparisons, in particular those made using the χ^2 test. In doing such comparisons, many assumptions must be made. For example, it is usually assumed that a term must appear in both corpora in order to be included in the analysis. This assumption ignores lexemes that are very specific to a particular corpus, and relaxing it produces different results. The differences are even more pronounced when the definition of "lexeme" is extended beyond individual words to bigrams, many of which are domain-specific. Results from various comparisons are presented, along with a suggestion for a new standard, text categorization, against which to compare the results.

1. Introduction

The desire to be able to compare two textual corpora is not a new one, and is shared across many academic fields. Computational linguists, such as those working in speech recognition, would like to know how much their language models must change to accommodate different corpora. Literary theorists would like to be able to determine the authors of anonymous texts. Sociolinguists would like to know the differences between language varieties, and which features are most characteristic of a particular variety.

Despite these desires, there has been limited work in corpus comparisons from a statistical standpoint, but not for lack of a good reason. Such comparisons are very difficult, as language is multifaceted. For example, two different sources might share the same frequency of personal pronouns, but differ significantly in the proportion of first, second, and third person pronouns. Biber (1994, 1995) introduces a promising approach to comparing different language varieties called "multi-dimensional analysis". The method involves counting the occurrences of *n* linguistic features in a corpus and then performing a factor analysis along these *n* dimensions. The first seven factors are selected for interpretation, and corpora are compared for differences along these factors.

Biber's approach, however, relies on a pre-existing linguistic framework and an assumption that each language variety selected by the researcher is made up of a homogeneous set of texts. Kilgarriff and Rose

* Many thanks to Jenny Staab, Doug Roland, and Ayşe Pinar Saygın for providing comments on an initial draft of this paper.

(1998) take a simpler approach with a set of highly informed statistical studies. They use word frequencies as data, and compare various statistics for their ability to measure "distance" between corpora. A very creative gold standard called "known similarity corpora" is employed to test the measures. They conclude that the χ^2 (chi-square) measure is superior to such measures as perplexity and the Spearman rank correlation coefficient. Kilgarriff (2001) expands upon these results and provides full details of the KSC method.

Other research on frequency-based corpus comparisons includes Hofland and Johannson (1982), and Leech and Fallon (1992). Both of these studies compared the American Brown corpus with the British Lancaster-Oslo/Bergen (LOB) corpus, the latter using the resulting data as a basis for comparison of the two cultures as they stood when the corpora were collected, in 1961. Rayson and Garside (2000) employ Dunning's (1993) log likehood measure to extract very salient features from air traffic control reports and ethnographic field notes. Holmes (1994) reports on many methodological problems in a seemingly simple examination of the evolution of sexist language use. Stubbs (1996) presents a few frequency-based studies in a book that explores this question: "How can an analysis of the patterns of words and grammar in a text contribute to an understanding of the meaning of the text" (p. 3)?

This paper can be viewed as a complement to Kilgarriff's (2001) comprehensive work that uses his methods under different assumptions with different corpora, and as an extension of Holmes (1994) that looks at more general problems from a statistical standpoint. The present work is not intended as a cure-all for the corpus comparison problem, but rather as a step in the direction of more principled, robust assumptions and tests.

The remainder of the paper is organized as follows: Section 2 introduces the corpora and methodologies, including the modification of the χ^2 test for corpus comparisons and methods for validating results. Section 3 presents three experiments that stretch the χ^2 test's capacity to accurately compare corpora. These experiments present only preliminary observations and data, so Section 4 discusses the results and suggests new directions for research on statistical corpus comparisons.

2. Method

2.1 The Corpora

The corpora used in this study come from six different Google Groups (formerly UseNet) discussion forums. Three are about computer-related topics; the other three are more general. Table 1 lists the group names and the abbreviations used to refer to each.

Table 1. The six corpora. The first three are collectively referred to as *computer corpora*; the next three are collectively referred to as *general corpora*.

Discussion Forum	**Abbreviation**
comp.lang	C-Language
comp.arch(itecture)	Comp-Arch
comp.graphics.algorithm	Graphics
rec.arts.books	Books
alt.atheism	Atheism
misc.consumers	Consumers

Each corpus is composed of 1.5 million tokens, with 150,000 tokens per year taken from messages in the years 1993-2002. *Token* is preferred to *word* (unless referring to a linguistic classification, such as *closed-class* word) because all alphanumeric strings, including numbers, acronyms, and proper names, are included in the analysis. Contractions are counted as single tokens, hyphenated words as two. The unit of analysis for the present studies, be it unigram or bigram, will be referred to as a *lexeme*.

2.2 Chi-square

Kilgarriff and Rose (1998) find the χ^2 test for statistical significance to be a good metric for frequency-based corpus comparisons, once it is adapted for that purpose. In its typical usage, the χ^2 test would determine whether a particular lexeme is drawn from the same underlying distribution in two different corpora. For most comparisons, thousands of lexemes will defeat the null hypothesis. Concerning this, Kilgarriff (2001) states: "This reveals a bald, obvious fact about language. Words are not selected at random" (p. 5).

Different corpora, however, will differ in their frequency of use of particular lexemes. Finding the average of these differences for popular (high frequency) lexemes, using the χ^2 test, yields a distance measure M between two corpora. M is simply the cumulative sum of the tests for individual lexemes divided by the total number of lexemes. A high value of M indicates great distance between two corpora. (*High* means high when compared with other values. The value of the measure, independent of any others, has no useful interpretation.) M is calculated as follows:

$$M = \frac{1}{n}\sum_{i=1}^{n}\frac{(O-E)^2}{E}$$

where n is the number of lexemes to be compared, O is the observed frequency of lexeme i in each of the two distinct corpora, and E is the expected value of the lexeme in each corpus. The expected value of lexeme i in corpus 1 is

$$E_{i,1} = \frac{N_1(O_{i,1} + O_{i,2})}{N_1 + N_2}$$

and likewise for corpus 2, substituting (in the numerator) N_2 for N_1, which are the sizes of corpora 2 and 1, respectively. In a case where the two corpora are of the same size, such as all of the ones investigated in the current study, the expected frequency is simply the average of the two observed frequencies.

This measure captures both style and substance, as closed-class words are not removed from the analysis. Here *style* refers to the closed-class function words such as pronouns, articles, and conjunctions that are typically excluded from statistical natural language processing tasks (Manning & Schütze, 1999), and *substance* refers to all other "content" words. If the lexeme you appears several hundred times more often in one corpus compared with another, as is the case with transcripts of telephone conversations and scientific papers respectively, this is significant to the measure.

2.3 Validating Results: Known Similarity Corpora and Text Categorization

How can one know whether the results of a corpus comparison are valid? This question is very familiar to machine learning researchers, who, for many tasks, require *gold standard* data. For example, if a classifier is to divide objects into categories, the objects must have predefined category labels. The validation methods of known similarity corpora and text categorization are described below.

2.3.1. Known Similarity Corpora

Kilgarriff (2001) presents a method called known similarity corpora (KSC) against which corpus comparison test results can be compared. In brief, KSC works by first collecting two corpora, *A* and *B*, that are known to come from different sources. In the experiments of Section 3, these corpora could consist of messages from two different newsgroups. Then new corpora are created: C_0, C_1, ... C_{10}, that are comprised of 100% documents from *A*, a 90/10% split between *A* and *B*, 80/20% *A* and *B*, etc.

This procedure yields a "distance" measure between corpora, at least in a rank-order sense. We know that C_0 is more similar to C_1 than it is to C_2, C_1 is more similar to C_2 than C_0 is to C_3, etc. A statistical measure of corpus similarity should be able to make these judgments like these.

This is the method that allowed Kilgarriff and Rose (1998) to conclude that the χ^2 test was superior to other competitors--its results were most closely aligned with the KSC rankings. KSC does suffer from a drawback that the authors acknowledge: it requires that the two corpora chosen for comparison are sufficiently similar that the most frequent lexemes in each have almost perfect overlap. Kilgarriff and Rose (1998) report that several of the statistical measures achieved 100% accuracy when the corpora were very different, and so could not be compared meaningfully with one another. As the experiments of Section 3 require comparisons between very different corpora, another potential gold standard is examined next.

2.3.2 *Text Categorization*

A more indirect method for creating a gold standard against which to judge statistical corpus comparisons is text categorization (TC). TC is the task of assigning text documents to one or more pre-defined categories. A machine learning classifier, for example one of the Bayesian probabilistic (Lewis & Ringuette, 1994), nearest neighbor (Yang, 1994), or decision tree (Lewis & Ringuette, 1994) variety is first trained on a set of documents with gold standard labels. Based upon a set of decision procedures that differs between classifiers, it then attempts to classify a test set of documents.

The simplest TC task involves a binary decision, wherein the classifier must place the document in one of two categories. One example is a spam filter; here the classifier must determine whether an e-mail message is spam or non-spam. The results reported below are also derived from a binary decision. The classifier's task is to determine which discussion forum generated a post, given a choice of only two that it has been trained on. Like KSC, this also provides a distance measure between corpora. The more difficulty the classifier experiences in the task (as measured by classification accuracy averaged over many trials), the more similar the two corpora are likely to be. The trivial case occurs when the documents in the two categories are drawn from the same source. In this case, the classifier can be expected to perform at chance (50%).

A naive Bayes classifier produced the accuracy results of the binary TC task found in Table 2. Each pair of corpora was trained and tested together. Full details of the training and testing procedures can be found in Appendix C.

Table 2. Pairwise percent accuracy measures for a naive Bayes classifier in categorizing documents in one of two categories. The matrix is symmetric.

	Graphics	**C-Language**	**Comp-Arch**	**Atheism**	**Consumers**	**Books**
Graphics	50.00	84.40	86.14	90.21	91.00	89.89
C-Language	84.40	50.00	84.67	88.37	90.92	88.41
Comp-Arch	86.14	84.67	50.00	91.29	89.68	89.57
Atheism	90.21	88.37	91.29	50.00	86.53	78.04
Consumers	91.00	90.92	89.68	86.53	50.00	84.69
Books	89.89	88.41	89.57	78.04	84.69	50.00

Table 3. Similarity ranks between corpora based on binary text categorization. The matrix need not be symmetric. For example, Consumers is most similar to Books, but Books is most similar to Atheism.

	Graphics	**C-Language**	**Comp-Arch**	**Atheism**	**Consumers**	**Books**
Graphics	1	2	3	5	6	4
C-Language	2	1	3	4	6	5
Comp-Arch	3	2	1	6	5	4
Atheism	5	4	6	1	3	2
Consumers	6	5	4	3	1	2
Books	5	4	6	2	3	1

Note the concentration of ranks 1-3 in the upper left and lower right quarters of Table 3, indicating the formation of two distinct groups. For even more clarity, Figure 1 shows a multidimensional scaling plot (Young & Hamer, 1987) of the numbers in Table 2.

This visual presentation helps to define what is meant by "distance" (or conversely, similarity) between corpora. While some of the relations between groups are not perfectly preserved when projected down to two dimensions (e.g., C-Language is actually more similar to Atheism than it is to Consumers), one observation is difficult to miss: as anticipated by Table 3, there are clearly two separate groups, which differ strongly along at least one dimension. They will hereafter be referred to as the *computer groups* and *general groups*.

While these are only initial steps in determining the suitability of TC as a gold standard for corpus comparison statistics, Figure 1 does make intuitive sense and provides a general trend that should stand out when using lexical frequency statistics to make corpus comparisons. The highly specialized, technical topics of conversation in the three computer related newsgroups are more likely to show lexical similarities to one another than to the more "diffuse" topics in the general groups, and the classifier exploits these tendencies.

TC is intended to provide a standard against which statistical measures of lexical frequency difference can be judged. Unlike KSC, which was designed specifically for this purpose, TC can provide a direct measure by itself. However, this is not of much practical value, as TC (in the form presented here) is several orders of magnitude slower than measures like the χ^2, Mann-Whitney, and *t* tests. Further discussion is provided in Section 4.2.

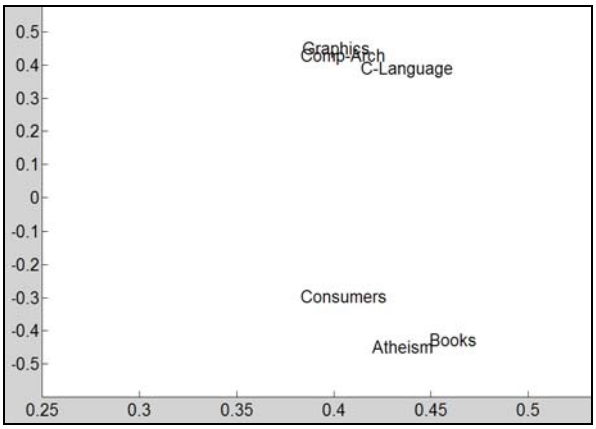

Figure 1. A multidimensional scaling plot of the data in Table 2. Two distinct groups of corpora emerge. An MDS of the rank data in Table 3 produces a qualitatively similar picture.

Table 4. Corpus-specific lexemes. Values are per 1.5 million tokens.

Corpus	Lexeme	Occurrences	Average in other five corpora
C-Language	*Goto*	446	7.4
Comp-Arch	*Caches*	550	3.8
Graphics	*coordinates*	879	2.4
Books	*Novels*	738	3.0
Atheism	*Theist*	449	0.4
Consumers	*Fees*	407	3.8

3. Corpus comparison experiments

3.1 Experiment 1: Minimum Frequency Assumption

Kilgarriff and Rose (1998) assume a minimum frequency of 5 observations per lexeme in each corpus, and use the top 500 most frequent lexemes in the joint corpus, which is created by combining the two corpora. This assumption is necessary, as the χ^2 test is known to be yield unreliable results when dealing with very low expected frequencies. (Snedecor and Cochran (1989) recommend a conservative expected frequency of greater than 20 for a lexeme if it is to be used in a χ^2 test.) For two sufficiently large corpora, intuition tells us that each of the 500 most common lexemes in the joint corpus will certainly occur at least 5 times in each corpus. With regard to the Google Groups corpora, this intuition is not correct.

A corpus that is specific to a domain may contain lexemes that occur far more frequently than in other corpora. Table 4 lists one of these lexemes for each corpus. For five of the six cases, the corpus-specific lexeme occurs with a frequency that is more than two orders of magnitude greater than its average frequency in the other five corpora.

Lexemes that occur zero times in one corpus and many times in another are valid, because the expected frequency is calculated using the joint corpus; a minimum observed frequency of 5 for any given lexeme in each corpus need not be assumed. Experiment 1 makes fifteen comparisons among the six corpora using the metric *M* defined above. The top 500 lexemes in each joint corpus are included in the computation. Figure 3 shows some of the comparisons of Experiment 1 for varying values of minimum frequency in each separate corpus.

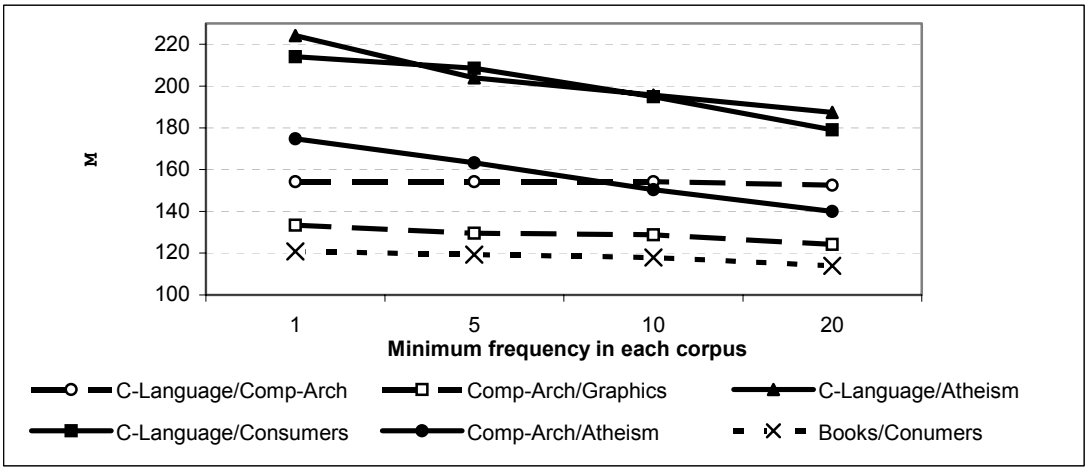

Figure 2. Corpus comparisons for different minimum frequencies of lexemes in each corpus. Some comparisons have been eliminated for clarity. In this and subsequent figures, dashed lines with open data points represent computer comparisons, short-dashed lines with + or × data points represent general comparisons, and solid lines with filled data points represent cross comparisons.

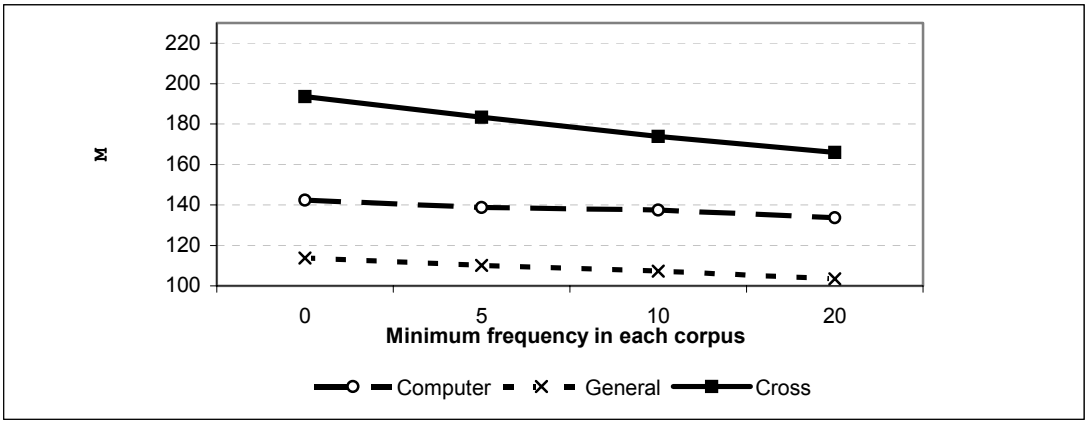

Figure 3. Average values of *M* for three types of comparisons for different minimum frequencies in each corpus. *Computer*, *General*, and *Cross* respectively refer to comparisons within the three computer groups, within the three general groups, and between these two types of group.

There are several things to notice about Figure 2. First, if minimum frequency could be arbitrarily selected, one would expect none of the lines in the graph to cross. Obviously, this is not the case. Some of the lines have a relatively flat slope, indicating that changing the minimum frequency in those particular comparisons had little effect on the resulting value of *M*. These lines belong to comparisons within the three general groups (hereafter called *general comparisons*) and within the three computer groups (hereafter called *computer comparisons*). The downward sloping lines belong to the comparisons between the general and computer groups (hereafter called *cross comparisons*).

The technical lexemes of the computer groups are shared enough that only a few have a frequency lower than 20 in one computer corpus yet high enough in the joint computer corpora to be included in the top 500. The same holds for general comparisons, which have fewer domain-specific lexemes and therefore share more high frequency lexemes that are fairly common to any corpus.

The cross comparisons, however, show that minimum frequency is not a parameter to be set arbitrarily. Many computer-specific lexemes occur fewer than five times in the general corpora, and so have an impact on *M* when the minimum frequency is changed. The most striking example of this is the comparison between Comp-Arch and Atheism. With a minimum frequency of 10 or 20 in each corpus, the Comp-Arch group would be deemed more similar to the Atheism group than to the C-Language group. This does not accord well with the similarity measures attained by the TC runs, but is understandable in light of the discussion above.

Figure 3 shows the average values of *M* for each of the three types of comparisons. The distance between comparisons in the cross category increases by 16.6% if, instead of assuming that a lexeme must appear at least 20 times in each corpus, it is assumed that it might not appear at all in one of the two corpora. This is contrasted with the computer comparisons (6.5%) and general comparisons (9.9%). This should put to rest any fears that the measure is only capturing trivial differences between corpora, such as the names of the discussants.

For the most part, using the χ^2 test and individual lexical frequencies is efficacious when compared against the TC standard as to which corpora should be more distant from one another. Recall that TC produced higher within-group similarities for both the computer and general groups than for comparisons across the two types of group. The χ^2 test does a better job of capturing this distinction when the minimum frequency threshold is reduced.

Figure 3 shows that, on average, cross comparisons are more affected by the choice of a minimum frequency for terms to be included in the calculation of *M*. However, the results are not very dramatic. This may be because individual unigrams are not the best terms to count when comparing two corpora. Experiment 2 addresses this.

3.2 Experiment 2: Bigrams

The analysis above indicates that individual term frequencies, isolated from any context, can only tell

us so much. Using bigrams, or adjacent tokens, as the units of analysis may provide a more appropriate test of distinction between corpora. For example, the terms *no* and *god* occur independently in many different contexts and corpora, but *no god* appears far more often in Atheism than in the other corpora under investigation

Table 5. Corpus-specific bigrams. Values are per 1.5 million tokens.

Corpus	Bigram	Occurrences	Average in other five corpora
C-Language	*void main*	257	2.0
Comp-Arch	*shared memory*	186	1.0
Graphics	*control points*	173	0
Books	*short stories*	193	0.4
Atheism	*no god*	234	1.6
Consumers	*credit report*	204	0

Damerau (1993) uses relative frequency ratios to identify domain-specific bigrams with good success in the task of genre classification. Kilgarriff and Rose (1998) exclude bigrams in their studies for several good reasons, one of which is KSC's requirement of "similar enough" corpora. A sample of corpus-specific bigrams, along with their frequencies, is shown in Table 3.

Experiment 2 is identical to the previous one, with the exceptions that frequency comparisons are made among bigrams, rather than unigrams, and the top 1000 most frequent bigrams in the joint corpus are used, rather than the top 500. This is because the number of possible bigrams is much greater, being the square of the number of unique unigrams. This parameter, referred to as n in the definition of M, is yet another that needs to be set[1]. Figure 4 shows some of the comparisons of Experiment 2.

The first point of interest in Figure 4 is the scale of M. It is much smaller than the scale of Figure 2. The values of M are now lower for two reasons. First, the top 1000 most frequent bigrams (n) are used, rather than the top 500; but the lower 500, despite being less frequent, still count with the same "weight" toward M because the χ^2 value that is produced is divided by n. Second, bigram frequencies are lower than unigram frequencies, owing to the much greater number of possible bigrams. The χ^2 test is affected this: as the number of observations decreases, the values it produces also decrease. Because of this, it must be emphasized once again that M is only to be used as a comparative measure, not an absolute measure.

Having said this, there are many interesting comparisons to make in Figure 4. To underscore the importance of using corpus-specific bigrams, note the dramatic difference between using values of 0 and 1 for the minimum frequency. In almost every case, this drop is proportionally greater than the drops between 1 and 5, and between 5 and 10. If it assumed that a bigram must appear in both corpora, then Books is judged to be equidistant from both Comp-Arch, a computer group, and Atheism, a general group, regardless of minimum frequency: 1, 5, or 10. However, if the 1000 most frequent bigrams in the joint corpus need not appear in both corpora, the results change dramatically, with Comp-Arch becoming more distant from Books than Atheism.

The same pattern holds for the general comparison between Books and Consumers and the cross comparison between Comp-Arch and Atheism. These two comparisons are judged to be roughly equal at the 1, 5, and 10 levels, but only at the 0 level are the TC data backed up: the general comparison is now much lower on the M scale than the cross comparison.

Finally, C-Language is compared with Atheism and Graphics, and these measures are found to slowly converge as the minimum frequency is increased. Their greatest separation comes at the 0 level. However, they do not immediately converge at the 1 level because, being computer corpora, C-Language and Graphics tend to share some corpus-specific bigrams in very low numbers that do not appear in Atheism.

Figure 5 shows the average values of M for each of the three types of comparisons.

[1] The effect of varying n is less than that of varying minimum frequency. This is probably because as n is increased, the new terms that are added have a decreasing impact on M. Such is the nature of the χ^2 test that the 500th most frequent term will not be nearly as important as the 50th most frequent term. See Experiment 3 for a more concrete example.

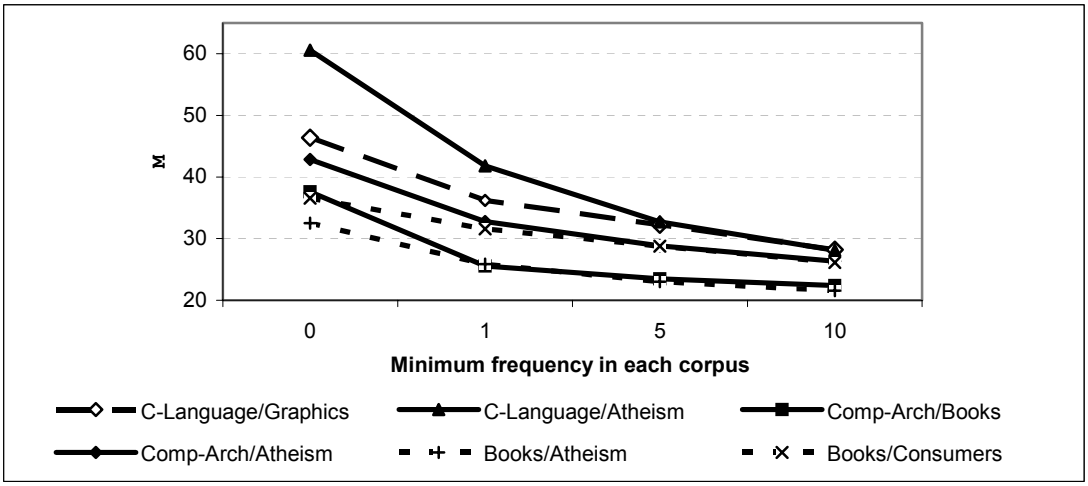

Figure 4. Corpus comparisons for different minimum frequencies of bigrams in each corpus. Some comparisons have been eliminated for clarity. Note that the minimum frequencies are different from those in Figure 2.

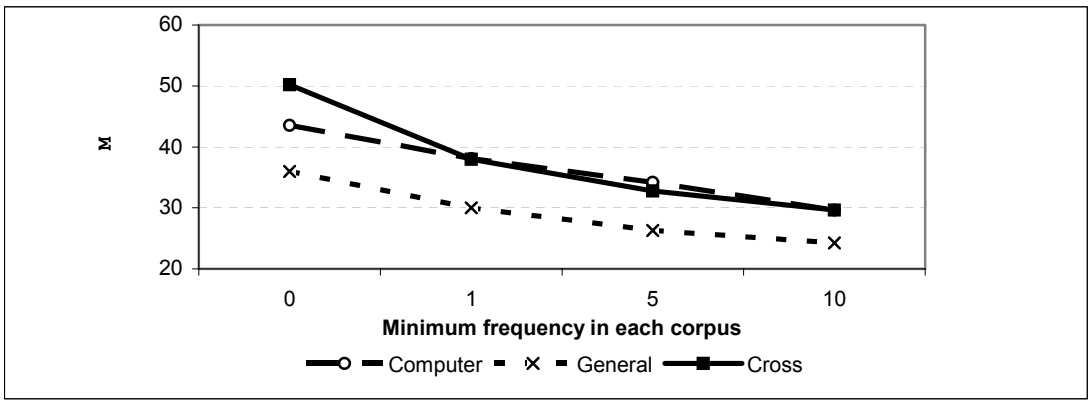

Figure 5. Comparison type averages for different minimum frequencies of bigrams in each corpus.

Figure 5 is notable for its difference from Figure 3. On average, if bigrams must appear in both corpora to be compared, then minimum frequency is not relevant to computer comparisons and cross comparisons: they are judged to be nearly identical at the 1, 5, and 10 levels. However, as was shown in Figure 4 for several of the individual comparisons, relaxing this restriction produces more reasonable results: the computer and cross comparisons are teased apart. When permitted to include very corpus-specific bigrams, distances within computer corpora are deemed closer than distances between computer corpora and general corpora. The percent increases in *M* between assuming that a bigram must occur *just once* in each corpus and the assumption that it need not occur in both corpora are: 32.2% for cross comparisons, 14.3% for computer comparisons, and 20.0% for general comparisons.

These results indicate that frequency-based comparisons of corpora using bigrams are probably more accurate than those using individual tokens, insofar as a measure of accuracy can be obtained--in this case, the TC data. But preliminary tests showed that more than half of the top 1000 bigrams in each joint corpus were composed of closed-class words, such as *in the*, *has not*, and *for a*. It is not clear whether including these bigrams helps or hinders the comparisons. Experiment 3 aims to address this question.

3.3 Experiment 3: Content-bearing bigrams

Should closed-class words be removed from the analysis? This is an issue of style vs. substance (see Section 2.2), and depends on one's reason for comparing corpora. For the task of predicting the next word in a sentence, it would seem more important to know that the word *space* frequently follows *address* in the computer architecture domain than to know that *the* follows *in* with slightly lower frequency when compared to other domains. If, however, the task is to identify an anonymous author, information about the frequencies of all words and bigrams is important, as different writers vary greatly in their frequency of use of different closed-class words and phrases (Craig, 1999).

What about *free will*? Words may have multiple senses that vary according to context. This feature of language known as *polysemy* has given birth to the field of word sense disambiguation (WSD; see Ide & Veronis, 1998) and many bad puns. Here, *will* is used in the *contentful* sense to mean a desire or purpose, not to indicate futurity in the closed-class sense.

There are many other examples, such as *no god* (for Atheism) and *pointer to* (for C-Language), that are important in characterizing a corpus but which contain closed-class terms. Bigrams that contain at least one *content* (open-class) *lexeme* will be referred to as *content-bearing bigrams*. They may, of course, contain two content lexeme.

Experiment 3 uses the same settings as Experiment 2, but considers only content-bearing bigrams in the analysis. Figure 6 shows some of the results. Removing all bigrams composed of two closed-class words appears to have little effect on the comparisons as a whole, producing results that are qualitatively very similar to those in Experiment 2. With a little reflection, this should not be surprising. Given these results, closed-class bigrams can be assumed to occur with roughly equal frequency in any of the Google Groups corpora. So, in removing them, the absolute numbers produced by the metric *M* may shift, but for the most part they will shift by roughly the same proportion for each comparison, because the element that was removed was not very important to the analysis in the first place. In performing corpus comparisons using bigrams, substance trumps style. Figure 7 shows this to be true in its similarity to Figure 5.

When using only content-bearing bigrams in the calculation of *M*, minimum frequencies of 1, 5, and 10 yield a value of the average cross comparison that is smaller than that of the average computer comparison. Here, only when the minimum frequency of a lexeme in each corpus is allowed to drop to 0 are the relative similarities found by TC upheld in *M*. This is, in part, due to the fact that less frequent bigrams are now included in the analysis (though many of them are still corpus-specific), whereas in Experiment 2 they would not have ranked in the top 1000 of a joint corpus drawn from a computer group and a general group. The percentage increases in *M* between the 0 and 1 levels are: 47.3% for cross comparisons, 18.4% for computer comparisons, and 33.5% for general comparisons.

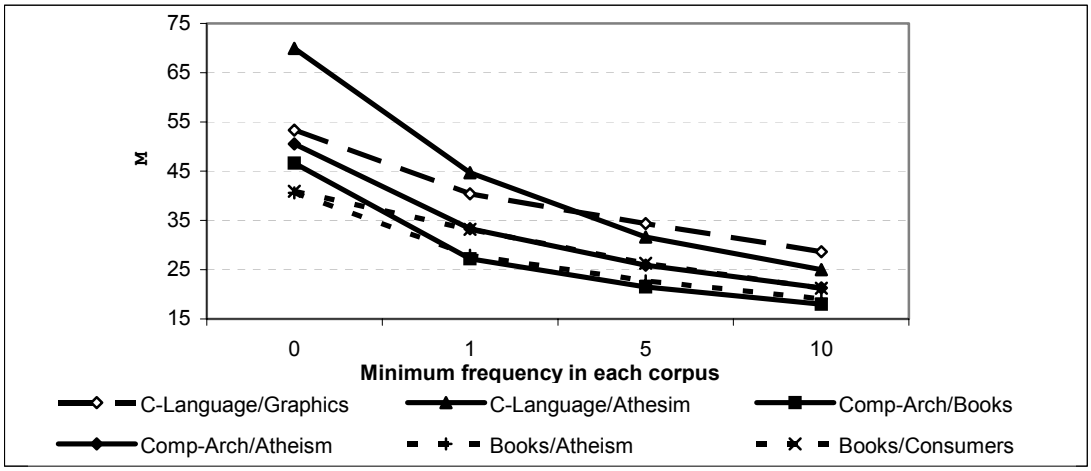

Figure 6. Corpus comparisons for different minimum frequencies of content-bearing bigrams in each corpus. Some comparisons have been eliminated for clarity.

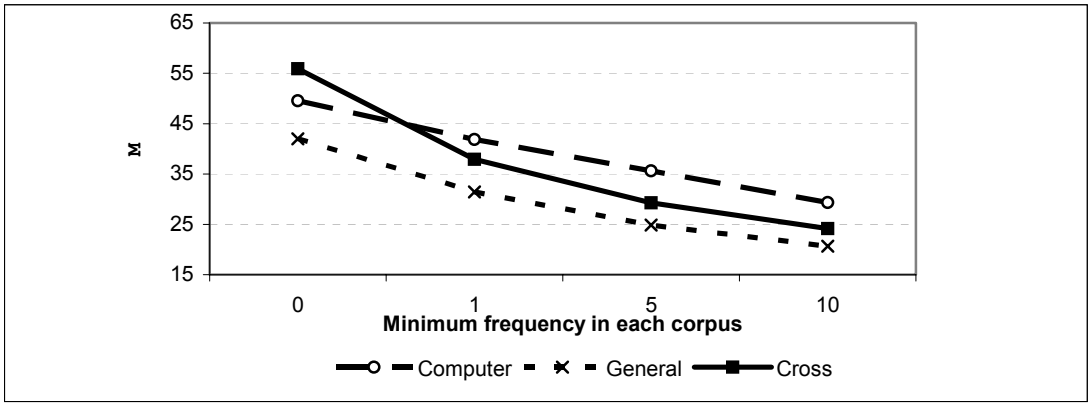

Figure 7. Comparison type averages for different minimum frequencies of content-bearing bigrams in each corpus.

4. Summary and Future Work

This study introduced new corpora, new testing conditions, and a new gold standard for the problem of statistical corpus comparisons. Many of the results are tentative and speculative, but suggest a rethinking of how to conduct corpus comparisons, and how to validate the results.

4.1 Discussion

Corpus comparisons based upon lexical frequency are useful up to a certain point. Certainly, if one wishes to pick out lexemes that are "significantly American" or "significantly British", as Leech and Fallon (1992) did in comparing the Brown and LOB corpora, frequency metrics are valuable. But when it comes to making similarity judgments between entire corpora, lexical frequency may be too crude a measure to capture an enriched notion of *similarity*. There are many linguistic dimensions along which corpora can be compared. Biber (1994, 1995) aims to make up for this crudeness, but does so at the price of couching his analysis in terms of pre-existing linguistic theory, and still must make choices as to which language features should be included in the analysis.

The present work has shown that small changes in initial parameters and assumptions can lead to significant discrepancies in the values assigned by a cumulative χ^2 metric, which Kilgarriff and Rose (1998) deem the best metric for frequency-based corpus comparisons among those tested. Even in a seemingly simple frequency-based study, many questions must be answered before undertaking a comparison of corpora. What value should be used for the minimum expected frequency of lexemes? How many unique lexemes should be included? Should closed-class words be removed? Should words be lemmatized? Kilgarriff (2001) provides some well-reasoned criteria on which to base answers some of these questions, but there are many more that can be asked.

Are frequency-based corpus comparisons so assumption-laden as to be useless? Certainly not. As has been echoed throughout this paper, this question ultimately depends on the task at hand. For a speech recognition system, where one of the goals is to reduce perplexity (roughly, the number of terms that are likely to immediately follow a given term), then the bigram measures presented above are useful, given the fact that some bigrams occur very frequently within particular topic domains and very infrequently in others. Discovering these domain-specific bigrams and incorporating them into a language model trained on one corpus can reduce the time necessary to port it to another corpus. As far as a task like author identification is concerned, the *bag-of-words* approach presented in this paper is not sufficient. More low-level stylistic analysis, such as has been advocated in the social science tradition for decades, must accompany any frequency measures.

4.2 Suggestions for Future Research

One of the issues raised in Section 2.3 was how to determine a *gold standard* against which to compare the results of corpus comparisons. Two methods were presented: known similarity corpora and text categorization. One line of research might involve testing the suitability of TC as a gold standard for similarity judgments.

TC has an intuitive argument in its favor, this being that a classifier will have greater difficulty in categorizing the documents drawn from two lexically

similar sources than from two lexically different sources. Whether the accuracy measure in the binary TC task is meaningful quantitatively is unclear, but the multidimensional scaling plot of Figure 1 certainly *seems* to capture an underlying truth in the distributions of lexemes across different types of group.

One way to bolster these results is to run the same binary classification tests using classifiers other than the popular naive Bayes that was used in the present study. Qualitatively, the results should not change. While the numbers of Table 2 may change due to the ability of the classifier to perform the task, the rank ordering of group similarities shown in Table 3 should show very little change.

Another line of research would involve a systematic comparison between KSC and TC. KSC suffers from the requirement that the two corpora to be compared must be similar enough that their most frequent lexemes have a very high degree of overlap. TC does not suffer from this drawback, so the two corpora may be very different. However, TC is a rather oblique way of determining similarity. It is computationally more expensive than KSC, which is an important consideration when corpus sizes reach into the hundreds of millions in tokens. A necessary first step would simply examine the results of TC when tested with mixed corpora created by KSC. The highest accuracy should be achieved on an even split between the two corpora (50% of documents drawn from each).

Finally, as was shown in the three experiments of Section 3, relaxing KSC's requirement of "similar enough" corpora leads to varying results of the χ^2 test when used as a measure of corpus similarity. This suggests a reexamination of how other statistical methods (e.g. Spearman, Wilcoxon) perform when TC is used as a gold standard and the minimum lexical frequency is varied.

References

Biber, D. (1994). An analytical framework for register studies. In D. Biber. & Finnegan, E. (eds.), *Sociolinguistic perspectives on register*. Oxford University Press.

Biber, D. (1995). *Dimensions of register variation.* Cambridge University Press.

Craig, H. (1999). Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them? *Literary and Linguistic Computing, 14,* 103-113.

Damerau, F.J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information processing & management, 29,* 433-447.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics, 19*, 61-74.

Hofland, K. & Johansson, S. (1982). *Word frequencies in British and American English.* Bergen: Norwegian Computer Center for the Humanities.

Holmes, J. (1994). Inferring language change from computer corpora: Some methodological problems. *ICAME Journal, 18,* 26-40.

Ide, N. & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational linguistics, 24,* 1-40.

Kilgarriff, A. (2001). *Comparing corpora. Manuscript*, ITRI, University of Brighton.

Kilgarriff, A. & Rose, T. (1998). *Measures for corpus similarity and homogeneity*. Manuscript, ITRI, University of Brighton.

Leech, G. & Fallon, R. (1992). Computer corpora: What do they tell us about culture*? ICAME Journal, 16,* 29-50.

Lewis, D. D. & Ringuette, M. (1994). Comparison of two learning algorithms for text categorization. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94).*

Manning, C. & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora, 38th annual meeting of the Association for Computational Linguistics (ACL 2000),* 1-6.

Snedecor, G.W. & Cochran, W.G. (1989). *Statistical methods.* Ames: Iowa State University Press.

Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture.* Oxford: Blackwell Publishers.

Yang, Y. (1994). Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proceedings of the 7th Annual International Conference on Research and Development in Information Retrieval (SIGIR 1994),* 13–22.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval, 1,* 69-90.

Young, F.W. & Hamer, R.M. (1987). *Multidimensional scaling: History, theory and applications.* Hillsdale, NJ: Erlbaum

Appendix A: Criteria for Selection of Discussion Forums

One reason that the Google Groups corpora are so appealing is that, unlike other corpora that are widely used in statistical natural language processing studies (e.g. Wall Street Journal), Google Groups is not the work of a few select writers (and ultimately, an even smaller group of editors). Rather, it represents a group with members that are trying to make themselves understood using their own idiolects, and so a group consensus must be formed regarding how to communicate. This stands in contrast to the typical newspaper corpus, in which a few writers try to make themselves understood by thousands of readers who (other than through letters to the editor, which are filtered by the editor) are not given an opportunity to respond and create a dialogue. The following criteria were used to select the discussion forums used in this study:

1. On average, the group must contain at least 3 new threads per day throughout the decade-long span of the corpus. Anything lower than this indicates a sparse discourse community.

2. In a given month, at least 30 different people must post to the group. This is because we want an accurate representation of an actual discourse community. Absence of a variety of writers in a group is frequently indicative of childish "flame wars" carried on by a very small community that can last for years. Flame wars, however, are an integral part of the language use on some groups, such as alt.atheism. This group was carefully examined to determine that, while there may be ad hominem attacks going on, the majority of posts are intended to make a point using original content, and that there are many discussants participating.

3. Moderated groups are not included because the moderator acts like a newspaper or magazine editor in screening the incoming messages. Ultimately, it is his decision to allow or disallow a post.

Appendix B: Criteria for Selection of Threads

Most discussion threads are taken from June and July to ensure that the discussions take place within a single year. Rarely does a thread persist for more than a few months. The following criteria govern selection of individual threads:

1. The thread must contain at least 3 messages. Threads with 1 or 2 messages frequently consist of advertisements and/or messages posted to the inappropriate group.

2. Messages must be posted to at most two different groups, and the two should be linked by the topic. For example, a discussion of Aristotle's Ethics might simultaneously take place on alt.philosophy and rec.arts.books. Messages that are "cross-posted" to more than two groups are often considered spam and have little relevance to at least one of the groups. Occasionally, if someone believes that another group would be interested in a particular message or thread, she posts to that group in addition to the groups in which the thread originated. These individual messages are kept, but if the thread then continues in three or more groups, the subsequent messages are discarded. Threads that are posted between two groups used in this study are not selected. In general, discussions take place within a single group.

Appendix C: Training and Testing in the Binary Text Categorization Task

For pre-processing of the documents, considerable effort went into removing both "quoted" text from previous messages and "signatures" that frequently appear at the bottom of messages. All header information was also removed. A stoplist was not used, nor was any lemmatizing, to ensure that the lexemes the classifier was trained on were the same ones used in the corpus comparisons.

The classifier was a naive Bayes network with no embellishments (e.g. boosting). For each corpus, 10,000 documents were chosen randomly, 1,000 per year. The remaining documents were not used in training or testing. The vocabulary was pruned by requiring lexemes to appear a minimum of 5 times across a corpus. (This was done merely to speed up the training phase; leaving all vocabulary in produced nearly identical results for a few sample comparisons.)

For training, 100 documents were chosen randomly for each corpus from the sets of 10,000 documents. For testing, a different set of 100 documents were chosen for each corpus. This train/test cycle was performed 100 times, and the results were microaveraged.