Defeating the Homogeneity Assumption

Anne De Roeck, Avik Sarkar, Paul H. Garthwaite

Faculty of Maths and Computing – The Open University – Walton Hall – Milton Keynes – MK7 6AA, U.K.
{a.sarkar; a.deroeck; p.h.garthwaite}@open.ac.uk

Abstract

The statistical NLP and IR literatures tend to make a "homogeneity assumption" about the distribution of terms, either by adopting a "bag of words" model, or in their treatment of function words. In this paper we develop a notion of homogeneity detection to a level of statistical significance, and conduct a series of experiments on different datasets, to show that the homogeneity assumption does not generally hold. We show that it also does not hold for function words. Importantly, datasets and document collections are found not to be neutral with respect to the property of homogeneity, even for function words. The homogeneity assumption is defeated substantially even for collections known to contain similar documents, and more drastically for diverse collections. We conclude that it is statistically unreasonable to assume that word distribution within a corpus is homogeneous. Because homogeneity findings differ substantially between different collections, we argue for the use of homogeneity measures as a means of profiling datasets.

Keywords: homogeneity, term distribution, corpus profiling.

1. Introduction

It is common practise in some areas of statistical Natural Language Processing (NLP), and Information Retrieval (IR), to assume that terms in a document occur independently of each other. This gives rise to the well known "bag of words" model for text which, in spite of numerous drawbacks (see Franz, 1997), has been used extensively. One of the major reasons behind this is the success of the vector-space approach to IR. The model also makes the application of standard mathematical and statistical techniques very convenient. At the same time, it is widely accepted that the term independence assumption is wrong, and that words do not occur independently of each other. Though some information retrieval techniques are said to work precisely because language "mildly" defeats the assumption (Manning and Schütze, 2000), the actual extent to which the occurrence of terms depends on other terms is relatively unexplored.

Term independence is related to the notion of homogeneity in term distribution. When text is seen as a "bag of words", terms are expected to distribute evenly throughout documents. Yet they do not. There is a growing literature which investigates "burstiness" in the distribution of content words in documents - i.e. the fact that repeated occurrences of an informative word in a document tend to cluster together (eg Church, 1995). By and large, however, function words are ignored, or assumed to distribute evenly throughout text, to the point of becoming uninformative. Indeed, Katz (1995) develops a model for bursty distributions of "concept terms", and distinguishes between function words and content words on the grounds that function words are distributed more homogeneously throughout text.

In short, the statistical NLP and IR literatures tends to make a "homogeneity assumption", either as a consequence of a "bag of words" model, or in the treatment of function words. In this paper, we show that the homogeneity assumption does not generally hold. In particular, we show that it does not hold for function words. Secondly, we show that datasets and document collections display different characteristics with respect to (non-)homogeneity, even when based on function words. Specifically, we show that the homogeneity assumption is defeated substantially for collections known to contain similar documents, and even more drastically for diverse collections. Evidence of homogeneity in term distribution rarely survives beyond very small text chunks.

We start from work in the corpus literature, which casts homogeneity as a property of term frequency distributions. Using Kilgariff's (1997) methodology as a base line, we use the χ^2 test (including the p-value) to relate a notion of homogeneity to a level of statistical significance. We explore different ways of partitioning datasets and investigate homogeneity in a range of collections with different characteristics. We also design, and report on, experiments where we investigate the effects of inducing different levels of randomness in text drawn from different collections.

Our results are significant in their own right, in that they show that it is statistically unreasonable to assume that word distribution within a corpus is homogeneous. In addition, in showing that document collections are not neutral with respect to the property of homogeneity, we make an indirect argument for using homogeneity measures to profile textual datasets. Such measures would help application developers to estimate the differences (including genre differences) between datasets. They would also help evaluation exercises, where experimental results can be supplemented with information on the characteristics of the dataset on which they were obtained.

2. Homogeneity

Homogeneity in document collections has been approached from different perspectives. We will concentrate on homogeneity as a property of term frequency distribution, or word count. Kilgariff (1997) describes a basic method for using measures of similarity to gauge homogeneity in a corpus. Starting from the position that no corpus can be more similar to another corpus than it is to itself, he casts homogeneity as internal similarity of distributions, between two halves of a document collection. Clearly, distributions of different features can be checked for similarity. Primarily interested in language variety, he proposes to measure term frequency distributions, and initially uses the χ^2 statistic. In the corpus literature, measuring this particular flavour of homogeneity has been linked to gauging the distance between corpora and genre detection. Rose and Haddock (1997) suggest using similarity based homogeneity measures to verify language model quality in corpus acquisition. Rayson and Garside (2000) show applications in the study of social differentiation in the use of English. Cavaglia (2002a) defines a homogeneous corpus as one that belongs to the same sub-language. In all these, focus tends to be on similarity as a means of establishing that two collections belong to the same genre or sub-language, by measuring lexical and syntactic features such as term frequency or POS distributions. A departure from this theme is Cavaglia (2002b), who uses term frequency and POS distributions together with unsupervised learning to generate corpora. Cavaglia (2001) uses homogeneity measures on web documents to judge the spread of documents based on certain keyword searches.

2.1. *How to measure homogeneity in term frequency distribution?*

Kilgariff (1997) sees homogeneity as internal similarity. His basic method for measuring homogeneity involves five steps:

(1) Divide the corpus into two halves by randomly placing text in one of two sub-corpora;

(2) Produce a word frequency list for each sub-corpus;

(3) Calculate the χ^2 statistic for the difference in term frequency distributions between the two sub-corpora;

(4) Normalise for corpus length;

(5) Iterate over successive random halves.

Kilgariff (1997) and Rose and Haddock (1997) partition their corpus by placing successive chunks of 5000 words in each half. The basic technique of comparing two halves of a corpus has been used with different similarity measures. Kilgariff (1997) adopts χ^2, Rose and Haddock (1997) and Rayson and Garside (2000) use G2. Alternatives include correlation on term rank frequency data, such as Mann-Whitney (Kilgariff, 1996) or Spearman's S (Rose and Haddock, 1997). Kilgariff and Rose (1998) compare Spearman's S with χ^2. Rayson and Garside (2000) deploy log-likelihood on different features, to expose different aspects of similarity. Cavaglia (2002b) uses relative entropy (Kullback-Leibler divergence), χ^2 and G2.

χ^2 is found to perform well in comparative experiments (Cavaglia, 2002b; Rose and Haddock 1997), as long as certain conditions are met. Notably, each of the individual frequency values must be greater than or equal to 5. Dunning (1993) states that most statistical tests assume some underlying distribution (usually either normal or Chi-Square (χ^2)). He shows through experiments that these assumptions can only be made if the sample size is large enough. He also discusses likelihood ratio tests and compares the results with that of Pearson's χ^2 test.

2.2. *The χ^2 Test and the χ^2 statistic*

At this point, it is important to clarify the relationship between the χ^2 test and the χ^2 statistic. The χ^2 test is a standard method to test the hypothesis that two or more samples are homogeneous, i.e. that they are drawn from the same population at random. In the SPlus software on a UNIX platform which we used for the experiments, the χ^2 test has three values in the output. First, the χ^2 statistic is calculated by the following formula:

$$\chi^2 = \Sigma ((O-E)^2/E)$$

where it tests the difference between expected (E) and observed (O) occurrences of events and is calculated with (N-1) degrees of freedom (this is the second output). N is the number of terms under consideration. The third output value is the p-value, a measure of confidence as to whether the two samples are statistically significant. The p-value is actually a probability depicting the level of confidence about the judgement based on the sample size. Being a probability, its value lies in the range 0 to 1. A value close to 0 indicates that, based on the sample size, the null hypothesis of similarity between two samples should be rejected.

The χ^2 statistic, on the other hand, has also been seen as a similarity measurement. In the case of perfect similarity (i.e. homogeneity in our case), one would expect the observed and expected occurrences to be close. Hence a lower χ^2 value would indicate greater similarity as compared to a higher χ^2 value. As a consequence, the χ^2 value may be viewed as a measure for comparing the similarity of two corpora, provided the degrees of freedom (N-1) is kept constant. This is due to the fact that a χ^2 value is calculated by summation over all the terms

under consideration, which leads to a higher value if more terms are considered. The effect of number of terms considered can be approximately nullified by dividing the χ^2 value by the degrees of freedom (N-1). The measure is called Chi-square By Degrees of Freedom (or *CBDF*). This is the corpus homogeneity measure used by Kilgariff (1997). Most other work (Kilgariff, 1996 and 1997; Rayson and Garside, 2000; Rose and Haddock, 1997) on corpus homogeneity also deploys the χ^2 statistic as a measure, rather than as a statistical test of significance.

Even a small departure from homogeneity can be detected if a sample's size is large enough, the p-value will get closer and closer to 0 as the sample size increases. One would like a measure of homogeneity that is not affected greatly by sample size, so that corpora of different lengths can be compared. Also, it is preferable if the similarity measure is compatible with a test of homogeneity: if two corpora are of similar size, the one with the larger value on the similarity scale should also have the smaller p-value for the test of homogeneity. Using CBDF as the similarity measure and the χ^2 test as the test of homogeneity gives these desirable properties.

3. Experimental Framework

3.1. *Homogeneity detection to a level of statistical significance*

Our aim of investigating the homogeneity assumption requires a more fine-grained tool than simple use of the χ^2 statistic as a homogeneity measure. We are interested in conditions under which non-homogeneity is detected, and in factors that affect the degree of non-homogeneity in different datasets.

We will adopt Kilgariff's outline methodology described in section 3.1, and conduct our experiments based on χ^2, because it is found to perform well in comparative experiments (Cavaglia 2002b; Rose and Haddock 1997), as long as certain conditions are met. (In particular, each of the individual frequency values must be greater than or equal to 5.) The settings in which we have conducted our experiments satisfy these criteria.

We will differentiate results in two ways by reporting the p-value as well as the CBDF statistic. Given a null hypothesis (in our case, homogeneity), the p-value allows us to estimate the strength of the evidence offered by the data. A p-value < 0.1 is usually interpreted as constituting weak evidence against the hypothesis, a p-value < 0.01 as strong evidence against, and $p< 0.001$ as very strong evidence against the hypothesis. Normally, a p-value < 0.05 is considered significant (moderate evidence against the hypothesis). In our case, a p-value < 0.05 will be taken to indicate that non-homogeneity is statistically significant. The CBDF measure relates to the text and indicates the level of heterogeneity.

3.2. *Now divide a corpus*

Kilgariff's basic method (section 3.1) requires a corpus to be split into two halves, by randomly placing text in one of two sub-corpora. The obvious question is how to execute this division? One way might be to dissolve document boundaries and split the corpus halfway. Kilgariff (1997) and Rose and Haddock (1997) dissolve document boundaries, but place consecutive chunks of 5000 words in each partition. Why chunks of size 5000 were chosen, rather than some other size, is not explained. The method of partitioning a document set raises important questions that may affect the outcome of similarity based experiments. A chunk size of 1, for example, would give randomness, which we would expect to see reflected in the experimental results. Also, can chunk size be chosen independently of the document sizes, or

genres, in a corpus? What are the implications for homogeneity experiments if chunks of varying sizes are considered?

To answer some of these questions, we experimented with alternative ways of partitioning a corpus, with different ways of handling document boundaries. We also investigated a range of smaller chunk sizes. Briefly, we conducted three experiments:

1. Choose a document and assign it at random to either of two partitions (docDiv experiment).
2. Divide each document in the middle, and randomly assign one half to either of the partitions, and the other half to the other partition (halfdocDiv experiment).
3. Remove document boundaries and repeat the same experiments of Kilgariff (1997) with various chunk sizes, from 5 to 5000, and observe the homogeneity measure (chunkDiv experiment).

Kilgariff (1997) measures homogeneity using all terms which occur more than 5 times in each of the partitions. Since the homogeneity measures we are deploying are based on word count, the inclusion of the most frequent terms means that the behaviour of function words will dominate the outcome of our experiments, and our measure of homogeneity is examining largely stylistic homogeneity.

To allow more detailed tracking of the distribution of very frequent terms, we will, for each experiment, report results at different values for N. Experimental results are shown in Tables 4 to 7. CBDF and p-values are averaged over iterations.

4. Datasets

We aim to investigate homogeneity in datasets with different characteristics, and considered corpora of various types and stylistic differences.

Data Set	Contents of the documents
AP	Copyrighted AP Newswire stories from 1989.
DOE	Short abstracts from the Department of Energy.
FR	Issues of the Federal Register (1989), reporting source actions by government agencies.
PAT	U.S. Patent Documents for the years 1983-1991.
SJM	Copyrighted stories from the San Jose Mercury News (1991).
WSJ	Copyrighted stories from the Wall Street Journal (1987-1989).
ZF	Information from the Computer Select disks for 1989/1990, copyrighted by Ziff-Davis Publishing Co.
OU	The Open University intranet and extranet web-pages.

Table 1. Description of content of each of the datasets

We selected the seven different datasets of the TIPSTER collection. Apart from availability, and its use as an evaluation and benchmarking standard, this collection has other advantages for our purposes. Table 1 lists the datasets and shows they are artificially compiled, with some drawn from a narrow base of similar text types, or from a particular domain. To contrast our results, we also experimented on data collected from the Open University Intranet. This data-

set is more diverse in terms of document type and domain content than the TIPSTER ones. Table 2 gives some basic profiling statistics, which show some of the bias in the datasets. DOE, for example, appears relatively uniform regarding text length, whereas FR shows the largest range. Comparing the ratio of new to old words gives an indication of domain diversity. There is a significant difference between the rate of new terms occurring, between the OU dataset (1 in 131 words) and the SJM dataset (1 in 260 words), in spite of their similar size. The WSJ and SJM sets are quite close in size and characteristics as well as in genre type, so we would expect them to behave in similar ways. Note also that the 10 most frequent terms of all TIPSTER collections are function words, but not in the OU dataset (Table 3).

Data Set	No of Docs	Corpus Length (words)	Average Doc Length (words)	No of Distinct Terms	Average Distinct Terms per Doc	Shortest Doc (words)	Longest Doc.
AP	242,918	114,438,101	471.1	347,966	238.25	9	2,944
DOE	226,086	26,882,774	119.0	179,310	72.90	1	373
FR	45,820	62,805,175	1,370.70	157,313	292.65	2	387,476
PAT	6,711	32,151,785	4,790.91	146,943	653.05	73	74,964
SJM	90,257	39,546,073	438.15	178,571	223.60	21	10,393
WSJ	98,732	41,560,108	420.94	159,726	204.26	7	7,992
ZF	293,121	115,956,732	395.59	295,326	168.42	19	75,030
OU	53,681	39,807,404	744.36	304,468	219.87	1	15,430

Table 2. Basic profiling statistics of each of the datasets.

Data Set	10 Most Frequent Terms
AP	the, of, to, a, in, and, said, s, for, that.
DOE	the, of, and, in, a, to, is, for, with, are..
FR	the, of, to, and, a, in, for, or, that, be.
PAT	the, of, a, and, to, in, is, for, said, as.
SJM	the, a of, to, and, in, s, for, that, is.
WSJ	the, of, to, a, in, and, s, that, for, is.
ZF	the, m, p, and, to, of, a, in, is, for.
OU	the, of, to, a, and, j, in, k, is, report.

Table 3. 10 Most frequent terms in each dataset

5. Experimental results

5.1. docDiv

The docDiv experiment maintains document boundaries and compares similarity of the two halves after assigning whole documents randomly to either partition. This experiment investigates homogeneity across documents in a collection. As Table 4 (and Figure 1a) shows, the

experiment finds non-homogeneity ($p<0.05$) in almost all cases. The exceptions are the AP and the DOE datasets when the 10 and 20 most frequent terms are used, and the WSJ and SJM datasets for the 10 most frequent terms. All the other datasets show statistical significance, with p-values of 0 or close to it (very strong evidence against the homogeneity hypothesis). CBDF values provide further insight into the corpus. In most cases, they are quite large, indicating high levels of non-homogeneity.

5.2. halfdocDiv

The halfdocDiv experiment induces a level of randomness in the individual documents, by dividing each of the documents exactly halfway and assigning each half to one of the partitions. This experiment is sensitive to homogeneity within documents. Again, there was evidence of non-homogeneity between the two partitions.

Dataset	**Number of Terms (N)**							
	10	**20**	**50**	**100**	**500**	**1000**	**7000**	**20000**
AP	**2.107** **0.1216**	**1.576** **0.2139**	2.583 0.0003	2.290 0	2.732 0	2.601 0	2.441 0	2.435 0
DOE	**1.172** **0.463**	**1.450** **0.160**	1.755 0.0259	1.983 0	1.838 0	1.786 0	1.795 0	1.872 0
FR	54.524 0	41.715 0	72.093 0	66.787 0	51.387 0	61.266 0	39.043 0	23.534 0
PAT	21.074 0	29.315 0	62.494 0	55.353 0	50.265 0	44.824 0	32.056 0	22.468 0
SJM	**3.595** **0.1193**	2.768 0.0077	3.231 0	2.976 0	3.012 0	2.959 0	2.560 0	2.511 0
WSJ	**2.358** **0.178**	2.663 0.0019	2.364 0	2.335 0	2.623 0	2.749 0	2.831 0	2.917 0
ZF	11.947 0	8.133 0	6.907 0	6.576 0	6.122 0	5.634 0	4.595 0	4.576 0
OU	232.913 0	158.520 0	94.749 0	67.293 0	32.663 0	25.181 0	14.224 0	8.297 0

Table 4. docDiv Results. Average CBDF and p-value for a dataset using the N most frequent terms. Values in bold indicate cases where the homogeneity assumption has not been defeated (p>0.05).

Dataset	Number of Terms (N)							
	10	**20**	**50**	**100**	**500**	**1000**	**7000**	**20000**
AP	**1.774** **0.087**	**1.473** **0.117**	**1.369** **0.057**	**1.271** **0.066**	1.171 0.021	1.187 0.0001	1.147 0	1.136 0
DOE	**0.728** **0.655**	**0.931** **0.533**	**1.054** **0.438**	**1.043** **0.372**	**1.061** **0.195**	**1.027** **0.285**	**1.014** **0.271**	**1.01** **0.182**
FR	7.905 0.001	9.549 0	11.627 0	11.642 0	8.847 0	8.166 0	6.543 0	5.336 0
PAT	20.360 0	15.568 0	16.017 0	11.886 0	7.694 0	6.243 0	5.102 0	4.611 0
SJM	**1.323** **0.3860**	**1.569** **0.3919**	**1.320** **0.4436**	**1.469** **0.1069**	1.332 0	1.297 0	1.240 0	1.242 0
WSJ	**1.563** **0.279**	**1.618** **0.248**	**1.342** **0.203**	**1.298** **0.260**	1.236 0.017	1.210 0.0007	1.178 0	1.150 0
ZF	**1.948** **0.1288**	**1.858** **0.116**	1.709 0.0283	1.609 0.0240	1.559 0	1.598 0	1.536 0	1.556 0
OU	7.721 0.033	6.103 0.0025	8.091 0	8.216 0	6.366 0	5.502 0	4.223 0	3.087 0

Table 5. halfdocDiv results. Average CBDF and p-values for a dataset using the N most frequent terms. Values in bold indicate cases where the homogeneity assumption has not been defeated (p>0.05).

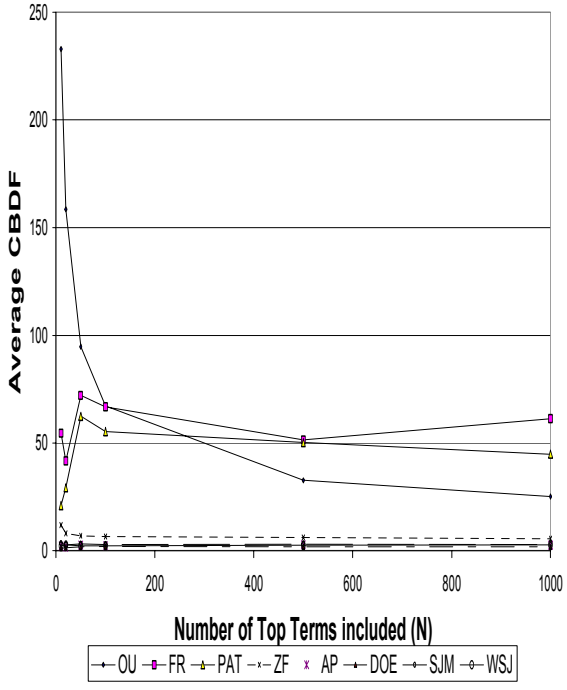

Figure 1a. docDiv

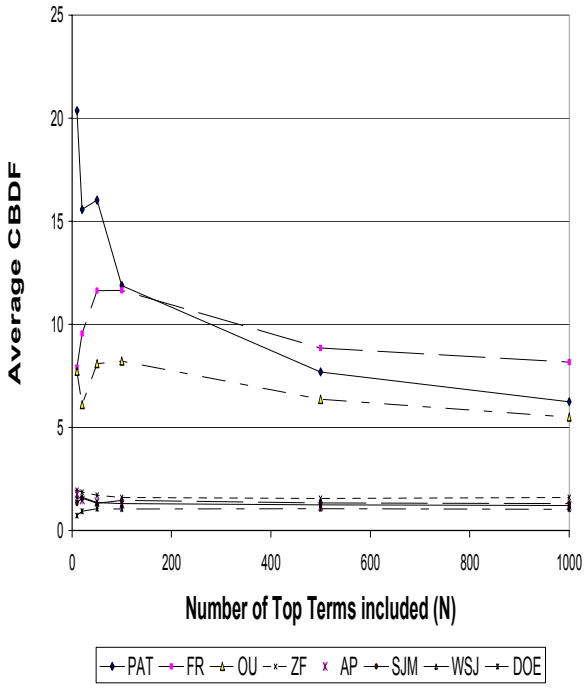

Figure 1b. halfdocDiv

Figure 1. (a) docDiv and (b) halfdocDiv for each dataset: Relationship between CBDF values and the N most frequent terms on which it is based.

Dataset	Number of Terms (N)							
	10	20	50	100	500	1000	7000	20000
AP	**0.628**	**0.836**	**0.871**	**0.984**	**0.990**	**1.007**	**1.018**	**1.012**
	0.7516	**0.6375**	**0.677**	**0.484**	**0.535**	**0.523**	**0.1595**	**0.179**
DOE	**1.141**	**1.225**	**1.151**	**1.050**	**1.038**	**1.002**	**1.008**	**1.008**
	0.3946	**0.3461**	**0.2505**	**0.3540**	**0.4229**	**0.462**	**0.431**	**0.3667**
FR	**0.754**	**0.961**	**0.967**	**1.033**	**1.016**	**1.025**	**1.022**	**1.013**
	0.650	**0.504**	**0.54**	**0.405**	**0.4174**	**0.335**	**0.2281**	**0.211**
PAT	**1.284**	**1.457**	**1.255**	**1.153**	**1.051**	**1.007**	**1.008**	**1.020**
	0.2451	**0.091**	**0.2273**	**0.1862**	**0.226**	**0.429**	**0.330**	**0.077**
SJM	**1.204**	**1.175**	**1.226**	**1.127**	**0.979**	**1.004**	**1.012**	**1.010**
	0.429	**0.375**	**0.293**	**0.268**	**0.608**	**0.454**	**0.262**	**0.181**
WSJ	**0.834**	**1.008**	**0.778**	**0.924**	**0.957**	**0.984**	**1.000**	**1.01**
	0.573	**0.492**	**0.822**	**0.679**	**0.682**	**0.6202**	**0.498**	**0.252**
ZF	**0.861**	**0.791**	**0.939**	**0.913**	**0.994**	**1.012**	**1.007**	**1.016**
	0.5781	**0.704**	**0.636**	**0.703**	**0.525**	**0.394**	**0.393**	**0.1258**
OU	**1.242**	**1.257**	**1.165**	**1.023**	**1.081**	**1.054**	1.042	1.033
	0.3395	**0.271**	**0.234**	**0.424**	**0.118**	**0.142**	0.034	0.005

Table 6. chunkDiv results with chunk size 5. Average CBDF values and p-values for a dataset using the N most frequent terms. Values in bold indicate cases where non-homogeneity is not statistically significant (p>0.05).

However, the experiment finds statistically insignificant non-homogeneity (p>0.05) much more often than the earlier docDiv experiment, with p-values higher than 0.05 for certain instances in more than half the datasets (Table 5; Figure 1b). Note that the DOE collection contains very short documents, each unlikely to deal with more than one topic. Also, CBDF values are much lower here than in the corresponding docDiv table. This suggests that terms distribute more homogeneously within documents, than across document boundaries. At the same time, with the 10 most frequent terms, there was evidence of heterogeneity among half-document partitions for three of the eight corpora, showing that, in general, even very frequent terms cannot be assumed to be uniformly distributed within a document. Hence this measure of homogeneity may be used to detect term burstiness in documents.

5.3. chunkDiv

Imagine dividing a dataset by randomly assigning each consecutive word to one of two partitions. Such a division would give randomness in the partitioning, and destroy any evidence of dependencies between terms. In this case, we would expect our experiments to seldom register statistically relevant evidence of non-homogeneity (we confirmed this experimentally for these datasets). On the other hand, Kilgariff (1997) reports non-homogeneity in partitions assigning chunks of 5000 words, which give far less randomness. Two questions arise. For a particular dataset, how large must the chunks be before non-homogeneity in the distribution of terms is statistically significant (p<0.05)? Is this level dataset dependent?

The chunkDiv experiment is designed to investigate the effects of different levels of randomness in a partitioning. We merged each dataset into a single document as in Kilgariff (1997), but placed a series of smaller chunks in partitions, ranging from 1 to 5000. We report only on chunk sizes 5 (Table 6; Figure 2a) and 100 (Table 7; Figure 2b).

Dataset	Number of Terms (N)							
	10	**20**	**50**	**100**	**500**	**1000**	**7000**	**20000**
AP	**0.824**	**1.105**	**1.412**	1.607	1.471	1.372	1.3004	1.3026
	0.6023	**0.3560**	**0.0735**	0.0019	0	0	0	0
DOE	**1.102**	1.864	1.646	1.511	1.354	1.414	1.4013	1.424
	0.3937	0.0280	0.0231	0.0317	0.0299	0	0	0
FR	**1.006**	**1.441**	**1.608**	1.803	1.924	1.834	1.782	1.746
	0.5071	**0.229**	**0.076**	0.025	0	0	0	0
PAT	4.181	3.051	2.682	2.420	2.252	2.104	1.977	1.876
	0.0232	0.0025	0.0007	0	0	0	0	0
SJM	**0.995**	**1.117**	**1.146**	**1.180**	1.410	1.402	1.317	1.291
	0.4720	**0.3851**	**0.3203**	**0.2463**	0	0	0	0
WSJ	**1.112**	**1.213**	**1.198**	**1.230**	1.196	1.283	1.2902	1.319
	0.3741	**0.324**	**0.2426**	**0.0937**	0.0383	0	0	0
ZF	**1.576**	**1.283**	1.709	2.190	1.41	1.673	1.315	1.884
	0.4152	**0.366**	0.011	0	0	0	0	0
OU	6.231	5.657	4.870	4.278	3.310	2.733	2.261	1.865
	0.0004	0	0	0	0	0	0	0

Table 7. chunkDiv results with chunk size 100. Average CBDF values and p-values for a dataset using the N most frequent terms. Values in bold indicate cases where non-homogeneity is not statistically significant (p-value>0.05).

Our results show a systematic relationship between increasing chunk size and increasing non-homogeneity: in Table 7, there are fewer non-significant p-values than in Table 6, and the CBDF values are higher. There also appears to be a relationship between registering non-homogeneity and a combination of document length and diversity of domain coverage. Where a dataset contains many very short documents, even small chunks are likely to cross document boundaries. Where such collections also cover diverse domains, documents are more likely to contain a higher proportion of distinct terms for the same amount of text. This would explain why the OU data started registering non-homogeneity at smaller chunk sizes than the other collections, as it combines a high incidence of short documents with diverse domain coverage.

Where they are function words, high frequency terms require bigger chunk sizes before non-homogeneity is apparent, when compared to experiments with more (less frequent) terms. Also CBDF values are lower when only high frequency terms are considered. To some extent, these results confirm Kilgariff (1996) and Katz (1995) who anticipate that more frequent function words have more similar distributions among documents than less frequent (content)

terms. Importantly, however, there are clear differences between the behaviour of function words in different datasets of the TIPSTER collection. (Results for the OU dataset are consistent with the conjecture of Kilgariff and Katz, because the OU most frequent terms contain non-function words). In all, the chunkDiv experiment revealed that the distribution of function words is very different from the distribution of content words.

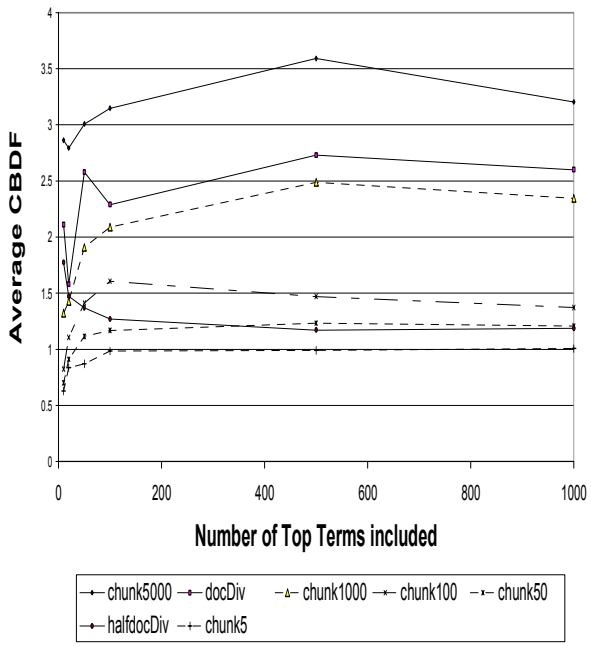

Figure 2a. AP Dataset

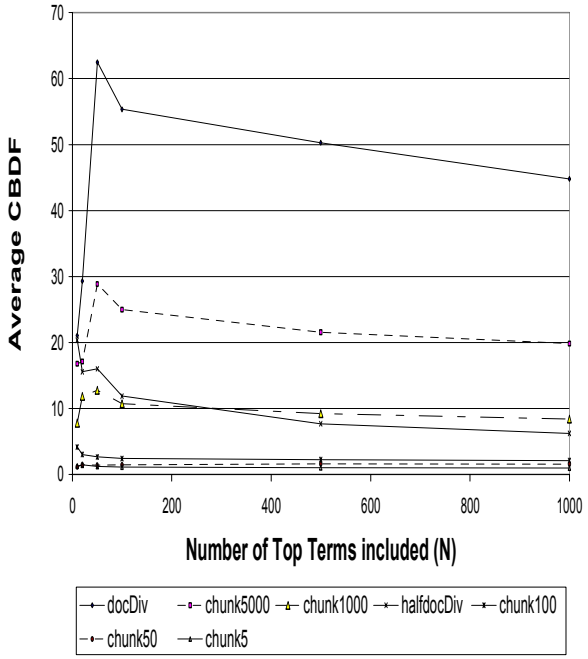

Figure 2b. PAT Dataset

Figure 2. (a) AP Dataset and (b) PAT Dataset. Relationship between CBDF values and N most frequent terms for all partitions (docDiv, halfdocDiv, and various chunk sizes).

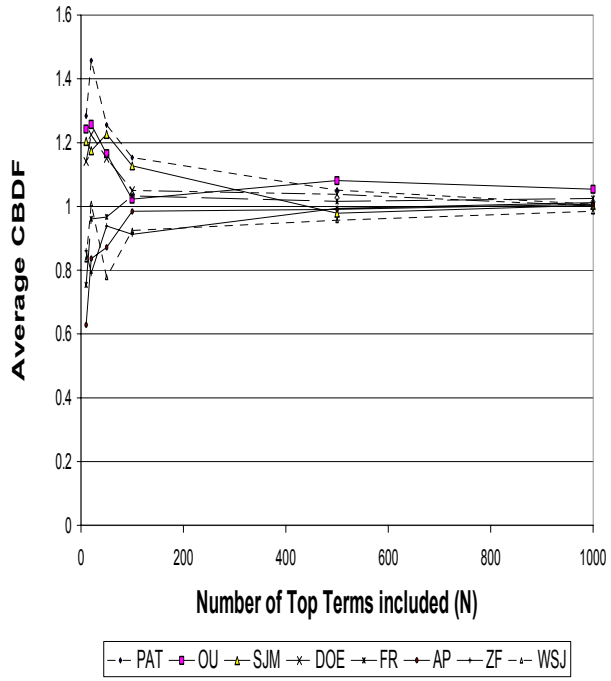

Figure 3a. chunksize 5

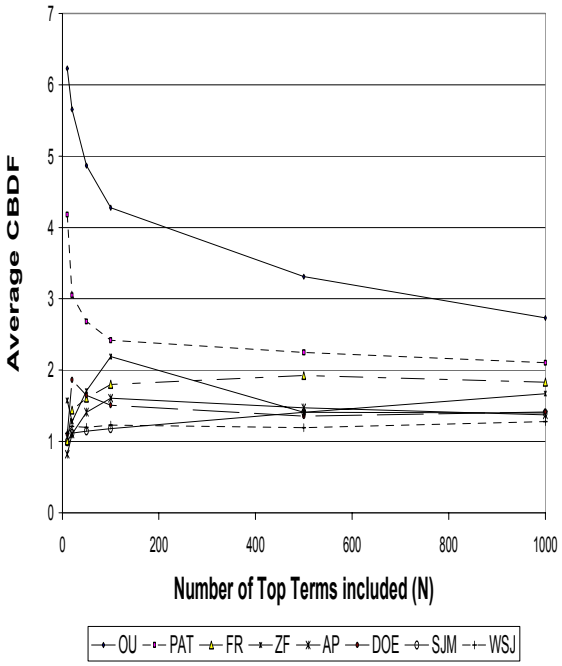

Figure 3b. chunksize 100

Figure 3 chunkDiv for (a) chunksize 5 and (b) chunksize 100 for each dataset: Relationship between CBDF values and the N most frequent terms on which it is based.

We plot average CBDF values of the docDiv. halfdocDiv and the various chunkDiv values for two datasets only, PAT and AP (Figure 2), due to lack of space. We show CBDF values using N most frequent terms up to N=1000. (There is not much variation for higher values of N.)

The figures show significant differences between the two datasets; the PAT dataset is much less homogeneous than the AP dataset. One may also note the increase in average CBDF value with increasing chunk sizes, which indicates increased non-homogeneity.

To show that collections have different homogeneity properties, we plot the results for all the datasets simultaneous for chunks of size 5 (Figure 3a) and size 100 (Figure 3b).

6. Contribution and Limitations

We have refined the Kilgariff (1997) method for measuring homogeneity by introducing a more fine grained approach for estimating homogeneity properties of datasets. We have supplemented the use of the CBDF measure with an indication of the strength of the evidence in the data. We have also experimented with dataset partitioning to investigate at which point non-homogeneity becomes detectable, and used chunk sizes together with the p-value to gain a view on how quickly the homogeneity assumption is defeated. Our approach has given us an opportunity to look closely at the behaviour of very frequent words and function words, which varies considerable across datasets. Our ultimate goal is to establish a collection of measures whose values might inform the deployment of techniques appropriate to the profile of the data. Whilst homogeneity measures can be used to estimate distance between datasets of different genres, much further work is needed to identify how the properties of datasets affect measures and values. Also, in the absence of results pertaining to less frequent (non-function) words, the work is of limited benefit for applications which make use of stop lists.

7. Conclusions and Future Work

We have investigated homogeneity in term distribution of the N most frequent terms. Starting from Kilgariff's work, we developed a notion of detecting non-homogeneity with a level of statistical significance, and have experimented with different partitions of a range of datasets. We conclude that the homogeneity hypothesis does not generally hold, even for function words. We also showed that different datasets will exhibit different homogeneity properties, which appear to correlate with a range of characteristics of the dataset. We conclude that it is statistically unreasonable to assume without question that word distribution within a corpus is homogeneous. In analysis of a corpus it is often very convenient to treat term distribution as homogeneous, and whether results would be biased to an important extent will depend on the analysis being performed and the purpose for which it is required. Our results show that it will also depend on the corpus being analysed, because the degree of non-homogeneity differs substantially between different collections. We argue for the use of homogeneity measures as a means of profiling datasets, in part to decide if an assumption of homogeneity is likely to lead to serious error.

Our objectives for future work are to find models of word distribution that fit reality better than the homogeneity assumption, for very frequent terms (including function words), and integrate them with "burstiness" phenomena for less frequent terms. Our chunkDiv experiments showed that the most frequent terms introduce a high degree of variability in homoge-

neity results, and we have started to investigate similar experiments where very frequent terms have been disregarded. We also want to investigate further the extent to which homogeneity measures are useful practical tools for profiling datasets.

References

Cavaglia G. and Kilgariff A. (2001). Corpora from the Web. In *Proceedings of the 4th Annual CLUK Colloquium*, Sheffield, UK, January 2001.

Cavaglia G. (2002a). Measuring the homogeneity of different varieties of language. In *Proceeding of the 5th National Colloquium for Computational Linguistics in the UK (CLUK)*, Leeds: 37-44.

Cavaglia G. (2002b). Measuring corpus homogeneity using a range of measures for inter-document distance. ITRI Report Series, ITRI-02-08, University of Brighton, UK.

Dunning T. (1993). Accurate Methods for the Statistics of Surprise and Co-incidence. *Computational Linguistics*, vol. (19/1): 61-74.

Franz A. (1997). Independence Assumptions considered harmful. In *Proceedings of ACL 1997*: 182-189.

Church K. (2000). Empirical Estimates of Adaptation: The chance of Two Noriega's is closer to p/2 than p^2. In *Proceedings of Coling*: 173-179.

Katz S. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, vol. (2/1): 15-59.

Kilgariff A. (1996). Which words are particularly characteristic of a text? A survey of statistical approaches. In *Proceedings AISB Workshop on Language Engineering for Document Analysis and Recognition*: 33-40.

Kilgariff A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings of ACL-SIGDAT Workshop on very large corpora*, Hong Kong.

Manning Chr. and Schuetze H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Rayson P. and Garside R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*, *ACL 38*, Hong Kong: 1-6.

Rose T. and Haddock N. (1997). The effects of corpus size and homogeneity on language model quality. In *Proceedings of the ACL-SIGDAT workshop on very large corpora*, Hong Kong: 178-191.