HEWLETT PACKARD

# The Effects of Corpus Size and Homogeneity on Language Model Quality

Tony Rose*, Nick Haddock
Interaction Technology Department
HP Laboratories Bristol
HPL-97-70
May, 1997

speech recognition, language modelling, email, dictation

Generic speech recognition systems typically use language models that are trained to cope with a broad variety of input. However, many recognition applications are more constrained, often to a specific topic or domain. In cases such as these, a knowledge of the particular topic can be used to advantage. This report describes the development of a number of techniques for augmenting domain-specific language models with data from a more general source.

Two investigations are discussed. The first concerns the problem of acquiring a suitable sample of the domain-specific language data from which to train the models. The issue here is essentially one of *quality*, since it is shown that not all domain-specific corpora are equal. Moreover, they can display significantly different characteristics that affect the quality of any language models built therefrom. These characteristics are defined using a number of statistical measures, and their significance for language modelling is discussed.

The second investigation concerns the empirical development and evaluation of a set of language models for the task of email speech-to-text dictation. The issue here is essentially one of *quantity*, since it is shown that effective language models can be built from very modestly sized corpora, providing the training data matches the target application. Evaluations show that a language model trained on only 2 million words can perform better than one trained on a corpus of over 100 times that size.

*Canon Research Centre Europe
Published and presented at the *Fifth Workshop on Very Large Corpora (WVLC5)*, University of Science and Technology, Hong Kong, 20th August, 1997

# The effects of corpus size and homogeneity on language model quality

*Tony G. Rose* [1] *(Canon Research Centre Europe) and*
*Nicholas J. Haddock (Hewlett-Packard Laboratories, Bristol)*

## Abstract

Generic speech recognition systems typically use language models that are trained to cope with a broad variety of input. However, many recognition applications are more constrained, often to a specific topic or domain. In cases such as these, a knowledge of the particular topic can be used to advantage. This report describes the development of a number of techniques for augmenting domain-specific language models with data from a more general source.

Two investigations are discussed. The first concerns the problem of acquiring a suitable sample of the domain-specific language data from which to train the models. The issue here is essentially one of *quality*, since it is shown that not all domain-specific corpora are equal. Moreover, they can display significantly different characteristics that affect the quality of any language models built therefrom. These characteristics are defined using a number of statistical measures, and their significance for language modelling is discussed.

The second investigation concerns the empirical development and evaluation of a set of language models for the task of email speech-to-text dictation. The issue here is essentially one of *quantity*, since it is shown that effective language models can be built from very modestly sized corpora, providing the training data matches the target application. Evaluations show that a language model trained on only 2 million words can perform better than one trained on a corpus of over 100 times that size.

## 1. Introduction

The development of robust speech recognition technology offers great potential for the design of improved interfaces to a wide range of applications. The current project concerns the development of one such application: the speech-to-text dictation of email messages. The work makes use of the Abbot recogniser, which is a connectionist/HMM continuous speech recognition system developed by the Connectionist Speech Group at Cambridge University. It is designed to recognise British English and American English, clearly spoken in a quiet acoustic environment (Hochberg et al, 1994).

The Abbot system is available with a vocabulary of 20,000 words, which means that anything spoken outside this vocabulary cannot be recognised (and therefore will be recognised as another word or string of words). The vocabulary and grammar (LM) were optimised for the task of reading from a North American Business newspaper, in this case the Wall Street Journal. Some 227 million words of training text were used in building this LM and it is widely used throughout the speech community. However, despite the size of the original training corpus, this LM was clearly not designed for the specific task of email dictation, so its performance is likely to be sub-optimal. However, a new vocabulary and LM can easily be created and then substituted for the one supplied.

## 2. Corpus Acquisition

### 2.1 The form of email messages

In order to build a LM for the task of email dictation, it is necessary to acquire a corpus of suitable email training data. However, behind this ostensibly simple objective lie several subtle challenges. "Email", as a general term, describes a great variety of types of communication. These types are perhaps best illustrated by considering the range of functions that email messages typically provide. For example, email can be used as a medium for:

---

[1] This work was completed at HP Labs during a previous appointment funded by the Royal Academy of Engineering.

- a formal face-to-face meeting;
- a casual face-to-face chat;
- a broadcast (e.g. "Tannoy") message;
- requesting information;
- replacing an office memo;
- replacing a phone call;

etc.

Clearly, the purpose of each communication can be very different, and the language used will reflect this. Furthermore, apart from the issue of domain (i.e. subject matter), the type of language used will also vary according to the social status of the participants. For example, when requesting advice from a mailing list one would tend to be more formal and polite than when requesting the same advice from a friend or colleague. Consequently, it would appear that email messages vary almost as much as spontaneous, spoken dialogue. If this is indeed so, then the prospects for building effective language models may appear somewhat limited. Clearly, in order to move forward, it is necessary to define some limits.

The first of these concerns the quantity. What is a reasonable size for an email corpus? There are few precedents for this so the question was answered empirically, by finding a compromise between the need to acquire sufficient training data and the need to complete the acquisition phase within a reasonable space of time. However, the time taken to reach a certain quantity depends very much on the source of the data, which forms the second limit. Simply put, from where should the email be acquired? A range of possibilities exist, e.g.: the Internet (i.e. bulletin boards, mailing lists, email archives) or specific individuals (i.e. previously saved messages, day-by-day output).

Evidently, email acquired from the Internet exhibits a wide range of authorship, function and subject matter. In addition, downloading large quantities of text from such sources without the authors' consent may involve certain copyright issues. Clearly, the limits of the source are more easily defined if the email is restricted to the output of a group of specific individuals. However, unless the group is very large, acquiring just 1 million words from their day-by-day output would be too slow to enable the acquisition to be completed within a reasonable space of time. Therefore, individuals with a large collection of previously saved messages were identified as more suitable candidates. Furthermore, a restriction that all members of this group must be employees of HPLB (Hewlett-Packard Labs, Bristol) placed a further constraint on the source. To ensure controlled authorship, only the outgoing messages of these individuals were collected.

## 2.2 The content of email messages

The "content" of an email message is not an easy concept to define. Evidently, the body contains much important content, but what about the other elements, e.g. headers, signatures, quoted sections, etc. - what role do they play? In the case of headers, a cursory analysis reveals that most can safely be discarded since they contain little useful information and as such should not be considered part of the training data. However, other email components are not quite so easily categorised. For example:

- quoted (included) messages: these are usually referred to by the message content, but are often the product of a different author;
- email 'signatures': these are often quite verbose, but rarely contribute anything to the message content;
- samples of Postscript/Latex: these were problematic, since people would often quote verbatim large passages to illustrate a point that did actually contribute to the content of the message. However, to build models from data that included such heavily marked-up text would be unwise.

Since the above items were all rendered as ASCII strings, their surface form could fairly reliably be predicted and they were therefore removed from the corpus using suitably designed "filters". However, there is a further number of email attachments that are not composed of predictable ASCII strings. These include items such as sharred or uuencoded files and word processor/DTP output. Evidently, such items need to be removed, but finding small fragments of such diverse data in a corpus of several million words is a non-trivial problem. Indeed, even after all the filters had been applied, it was often still necessary to manually inspect the results to ensure that the final corpus was as "clean" as possible.

Alternatively, rather than trying to filter out "noise" from the email "signal", it is possible to adopt the converse approach, and try to identify those lines which constitute genuine English within the overall email data, which may then be retained as the true training data. It is possible to achieve this using various heuristics, e.g. "retain those lines that contain at least 90% English words". However, this approach assumes there exists some

predefined vocabulary, which creates a tautology since the vocabulary is one of the things we seek to define in the first place.

### 2.3 The email collection process

A programme of email data collection took place over a period of 2-3 weeks, following the principles described above. This resulted in the acquisition of some 4 million words of email data. The "donors" were asked to provide both previously saved messages and intermittent day-by-day output. This was necessary since (by their very nature) saved messages tend to possess some sort of significant content, and were therefore often of above average length. In contrast, much day-by-day email correspondence uses an informal dialogue that is heavily context dependent, and therefore may be no more than a single brief phrase or sentence. This was then filtered in the manner described above. The final output was a corpus of email data of 1,962,280 words (i.e. 49% of the original size). This was then partitioned (95% : 5%) into training and test data.

## *3. Corpus augmentation*

It is theoretically possible to build a LM using the tiniest of corpora. On balance, however, the 2 million words of email training data looks somewhat inadequate compared to the 227 million words used for the WSJ LM. The problem is that the coverage of the n-grams is likely to be sparse, so any LM so built will be degenerate since it does not accurately predict the behaviour of the source. To illustrate, consider the distribution of unigram frequencies: a mere 14,137 word types (19%) in the email corpus have frequencies of 6 or greater. Therefore, to acquire a vocabulary of just 20k words (which is quite modest by current standards) without using frequencies of 5 or less clearly requires a training corpus larger than 2 million words.

It would be highly desirable therefore if a method could be devised whereby information from a large corpus could be combined with a smaller sample of the domain-specific training data to create an optimal language model. One such approach involves augmenting a base model built from a large corpus with information from a small sample of the domain-specific language (in this case email). There is evidence to suggest that this method can improve recognition performance (e.g. Rudnicky, 1995, Vergyri, 1995). An alternative approach is to use a suitable similarity metric to acquire further 'email-like' training data from a larger, more general corpus (henceforth referred to as the "remote corpus"), and then build a new language model from the combined text. This approach offers interesting possibilities regarding the development of a general methodology for corpus acquisition, since it should be possible to "grow" a suitable corpus of training data for any domain, using only a small sample as a "seed". A number ways to implement this technique have been developed, with varying degrees of sophistication and effectiveness. Broadly speaking, they fall into two categories: "top-down" methods and "bottom-up" methods.

### 3.1 The top-down approach

At its simplest, this approach involves a combination of manual inspection and regular expression searching to identify those parts of the remote corpus that contain suitable material. It relies on a good classification scheme and reliable organisation of the remote corpus. The British National Corpus (BNC) is a suitable example, since it contains 100 million words of modern English, both spoken and written, sampled from the widest range of materials. It is annotated with part-of-speech codes, and SGML-encoded according to the Text Encoding Initiative's Guidelines (Burnard, 1995). It is therefore possible to use the SGML tags to identify suitable texts. For example, extracting 10 million words of text for a domain such as World Affairs is trivially easy, since domain information is encoded in the header of each individual file (of which there are over 4,000).

Since much HP email concerns the computing business, and the BNC classifies computing as a branch of Applied Science, it would appear that the 10 million words from Applied Science section of the BNC may prove sufficiently similar. Likewise, the 10 million words classified as Commerce and Finance may also prove suitable. The effect of such an addition would be to increase the size of the training corpus from 2 million words to 22 million, which constitutes an increase of 1100%.

However, methods such as this cannot be justified by subjective judgement and anecdotal evidence. What is required is an objective measure that reliably identifies which of the domains in the BNC is most similar to HP email. There may be many standard statistical techniques for measuring the degree of similarity of two data sets, but not all are suitable for the task of comparing corpora (Church et al, 1991). For example, some assume a normal distribution, which is clearly inappropriate for textual data. What is needed therefore is a test that makes few assumptions about the distributions of the underlying data, but provides a directly usable measure of similarity. One such test is the rank correlation, using Spearman's S.

The assumptions behind rank correlation are few. It measures the degree of monotonic association between two rankable variables. The distribution of r as normal (mean 0, variance 1/(N-1), assuming independence) is asymptotic for large enough samples, and does not make any assumptions about normality. This test was therefore applied to the word frequency lists of each of the domains in the BNC and the email corpus, to identify which corpora were most similar. The correlation with the BNC as whole was also measured. All calculations used Spearman's S, where $D^2$ denotes the sum of the squares of the differences between the ranks of each pair, and N the number of ranked pairs:

$$r = 1 - \frac{6D^2}{N^3 - N}$$

It is known that a sublanguage corpus can have very different characteristics to a general corpus (Biber, 1993), yet it is not obvious how the position on this scale of a given corpus can be assessed. Consequently, it is necessary to determine the *homogeneity* of a corpus prior to performing any similarity measures, since it is not clear what a measure of similarity would mean if a homogeneous corpus was being compared with a heterogeneous one (Kilgarriff, 1996). A homogeneity test was therefore performed on the corpus of each domain. This was calculated using the following algorithm:

1. For each domain corpus, do (times 10)
    2.1 Divide the corpus into two halves, by randomly placing 5000 word chunks in one of two subcorpora;
    2.2 Produce a wfl for each subcorpus;
    2.3 Calculate the rank correlation between the two subcorpora;
3. Calculate the mean and standard deviation of r.

Table 1 shows both sets of results. The homogeneity values are across the diagonal, with mean and std. deviation shown in each cell. The other cells show the rank correlation (r) and the value of N. "BNC" refers to the complete corpus. The subdomains are labelled as follows:

| | |
|---|---|
| As: | Applied Science |
| Ar: | Arts |
| Be: | Beliefs & thought |
| Cf: | Commerce & Finance |
| Im: | Imaginative |
| Le: | Leisure |
| Np: | Natural & pure science |
| Ss: | Social science |
| Un: | Unclassified |
| Wa: | World affairs. |
| BNC: | The whole of the BNC |
| Email: | the 2 million word email corpus |

For large samples like these the rank correlation coefficient has a normal distribution with mean 0 and variance 1/(n-1) where n is the number of common words. Although the significance of the correlation is not in doubt, the differences are highly significant too. The difference between two rank correlation coefficients will be normally distributed with mean 0. The maximum possible value for the standard deviation is $(1/\sqrt{(n1-1)})+(1/\sqrt{(n2-1)})$ where n1,n2 are the two common vocabulary sizes. Any difference greater than about 0.03 is therefore significant, and there are many pairs for which this is true. It is therefore possible to rank the rank correlations, and hence the BNC domains.

Evidently, the strongest correlation with the email corpus is from As (Applied Science). Interestingly, this figure is higher than that between email and the whole BNC. The second highest domain correlation is with Cf (Commerce & Finance). This agrees with intuitions based on a manual inspection of the contents of the email corpus. The table also shows a polarity of the BNC - the "arts" domains at one pole, attracting each other (e.g. Ar:Bt = 0.408) but repelling the sciences (e.g. Im:As = 0.159). Similarly, the sciences attract each other (e.g. Np:As = 0.315). In the middle are domains such as World Affairs, Social Sciences & Commerce & Finance which correlate with both poles to varying degrees. Moreover, the Email corpus really stands out on its own, having a very poor correlation with the others (in many cases it is negative). This suggest that even if the most strongly correlated domains are chosen, it is difficult to justify augmenting the email corpus with texts selected from the BNC using this method.

| | Ar | As | Bt | Cf | Im | Le | Np | Ss | Un | Wa | BNC | Email |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ar | 0.789 | | | | | | | | | | | |
| | 0.001 | | | | | | | | | | | |
| As | 0.255 | 0.758 | | | | | | | | | | |
| | 72127 | 0.001 | | | | | | | | | | |
| Bt | 0.408 | 0.238 | 0.581 | | | | | | | | | |
| | 69932 | 73839 | 0.002 | | | | | | | | | |
| Cf | 0.340 | 0.351 | 0.291 | 0.730 | | | | | | | | |
| | 70648 | 70452 | 72970 | 0.001 | | | | | | | | |
| Im | 0.404 | 0.159 | 0.340 | 0.215 | 0.887 | | | | | | | |
| | 67604 | 73786 | 70807 | 72800 | 0.001 | | | | | | | |
| Le | 0.459 | 0.284 | 0.310 | 0.337 | 0.407 | 0.824 | | | | | | |
| | 67013 | 71281 | 71768 | 70407 | 67615 | 0.001 | | | | | | |
| Np | 0.122 | 0.315 | 0.150 | 0.161 | 0.061 | 0.145 | 0.662 | | | | | |
| | 75822 | 71596 | 76151 | 75339 | 76628 | 74952 | 0.002 | | | | | |
| Ss | 0.409 | 0.342 | 0.422 | 0.470 | 0.278 | 0.325 | 0.200 | 0.812 | | | | |
| | 68263 | 70405 | 69756 | 68023 | 70475 | 69796 | 74053 | 0.001 | | | | |
| Un | 0.273 | 0.161 | 0.174 | 0.229 | 0.244 | 0.357 | 0.032 | 0.226 | 0.486 | | | |
| | 74110 | 76754 | 77136 | 75423 | 74635 | 72373 | 79904 | 75161 | 0.002 | | | |
| Wa | 0.423 | 0.280 | 0.395 | 0.432 | 0.311 | 0.395 | 0.130 | 0.469 | 0.284 | 0.865 | | |
| | 68005 | 71298 | 70189 | 68404 | 69838 | 68296 | 75399 | 66921 | 73981 | 0.001 | | |
| BNC | 0.611 | 0.497 | 0.505 | 0.541 | 0.578 | 0.605 | 0.307 | 0.609 | 0.381 | 0.653 | 0.687 | |
| | 63522 | 66732 | 67920 | 66189 | 63755 | 63439 | 71615 | 63779 | 72033 | 62450 | 0.001 | |
| Email | 0.012 | 0.093 | -0.026 | 0.055 | -0.062 | -0.003 | -0.032 | 0.035 | -0.066 | -0.012 | 0.073 | 0.362 |
| | 81648 | 79745 | 82897 | 80884 | 83083 | 81908 | 82719 | 81143 | 84338 | 82015 | 80085 | 0.002 |

**Table 2. Similarity and homogeneity of BNC domains and email**

Table 1 also shows the results of the homogeneity tests. Email is by far the most heterogeneous, more so even than the "Unclassified" section of the BNC(!) This brings into question the results of the similarity measures in which the email corpus was involved, and mitigates further against the strategy of augmenting the email corpus with texts selected using the top-down method. Although, it does lend an important insight into the nature of the problem, since a LM derived from a heterogeneous corpus *should* have a higher perplexity than an equivalent one derived from a more homogenous corpus. However, homogeneity is a measure of unigram distributions, whereas perplexity is concerned with n-grams (where n is usually <=3), so it is not certain that the two measures would be directly related.

Clearly, the above results are highly revealing. However, they must be treated with some caution, since there are a number of limitations to the methodology. Most important of these is the fact that rank correlation compares differences in rank, ignoring absolute value (which can be significant). To illustrate, consider a case where the word "of" is ranked 3 in one corpus and 6 is another. This is a very important difference. Conversely, if "banana" is ranked 10,000 in one corpus and 100,000 in another, this is a very insignificant difference. But the difference of ranks for "of" = 3, for "banana" = 90,000. Clearly this technique is missing something important.

## 3.2 The bottom-up approach

The top-down approach assumes that the BNC classification system is perfect, in that each text classified as belonging to a certain domain really belongs in that domain. However, this is ultimately a subjective judgement, and frequently more than one classification is possible or even preferable (Lewis, 1992). Moreover, it is often the case that texts from the same medium are more similar to each other than texts from the same domain (e.g. a journal paper on computing may be more similar to a journal paper on geology than an item from a popular computing magazine, because the 'content' features are lost among the much more salient 'genre' features). Besides, no classification system is 100% reliable, so techniques that are based on them will inherit this uncertainty. Furthermore, domains such as Applied Science are very coarse-grained: they contain many more

types of material than just those of computing. Even if such corpora are subdivided to a further level of classification they still suffer the same problem, albeit at a finer level of detail.

An alternative strategy is to work in a "bottom-up" direction. In this approach, a similarity metric is used to find and extract related material from the remote corpus, regardless of their top-down classification. This method may not be as structured as the previous approach, but it is more robust in that it involves no manual intervention and does not rely on correct organisation or SGML tagging of the remote corpus. Moreover, it will not "miss" material that is classified under an unexpected domain or medium but is otherwise suitable. The success of this approach depends on the use of a reliable similarity metric (even more so than the top-down approach, since it is now being applied to each of the 4,000+ files in the BNC rather than the 10 domain-based collections). The rank correlation has been shown to have significant shortcomings, so it is unlikely to be optimal. An alternative statistical measure is the log-likelihood statistic.

The log-likelihood statistic, $G^2$, is a mathematically well-grounded and accurate method for calculating how "surprising" an event is (Dunning, 1993). This is true even when the event has only occurred once (as is often the case with linguistic phenomena). It is an effective measure for the determination of domain-specific terms (e.g. Daille, 1995) and can be also used as a measure of corpus similarity. In the case where two corpora are being compared, it is possible to calculate the $G^2$ statistic either for single words (as in the 2×2 table) or for a vocabulary of N words (as in an N×2 table). Using this statistic to find texts that are similar to email in the BNC could be achieved using the following algorithm:

1. Create a wfl for the email corpus.
2. For each individual text in the BNC, do:
    2.1 Create the wfl for the BNC text
    2.2 Create a contingency table from the 2 wfls (ignoring function words)
    2.3 Calculate the $G^2$ statistic and length of the contingency table
3. Store the filename, title, table size and $G^2$ in RESULTS
4. Output the RESULTS sorted on the $G^2$ statistic

Although the $G^2$ may be applied to the wfls regardless of their content, it was found empirically that performance improved if function words were excluded from the contingency table. A stop list of 241 function words was therefore applied in Step 4 of the above algorithm. The algorithm was run on the entire BNC (i.e. each of its 4,000+ files). The output was a sorted list of the files that most closely matched the wfl of the email corpus. The top and bottom 10 texts on this list are as follows:

```
/BNC/1.0/H/H4/H4L  MEDICAL CONSULTATIONS -- AN ELECTRONIC TRANSCRIPTION  23226    108.897
/BNC/1.0/G/G5/G54  MEDICAL CONSULTATIONS -- AN ELECTRONIC TRANSCRIPTION  23226    244.251
/BNC/1.0/H/H5/H58  MEDICAL CONSULTATIONS -- AN ELECTRONIC TRANSCRIPTION  23228    318.442
/BNC/1.0/F/F7/F78  STAFF MEETING -- AN ELECTRONIC TRANSCRIPTION           23226    358.087
/BNC/1.0/H/H5/H5B  MEDICAL CONSULTATIONS -- AN ELECTRONIC TRANSCRIPTION  23230    385.826
/BNC/1.0/G/G4/G4Y  MEDICAL CONSULTATIONS -- AN ELECTRONIC TRANSCRIPTION  23231    419.826
/BNC/1.0/K/KN/KNU  SPOKEN MATERIAL FROM RESPONDENT 716 -- AN ELECTRONIC  23227    425.806
/BNC/1.0/J/JJ/JJJ  BRISTOL UNIVERSITY -- AN ELECTRONIC TRANSCRIPTION     23231    526.359
/BNC/1.0/A/A9/A9B  GUARDIAN, ELECTRONIC EDITION OF 19891210; APPSCI MAT  23232    532.956
/BNC/1.0/H/HK/HKC  FREEMANS                                              23234    547.160
.......
/BNC/1.0/C/CB/CBG  TODAY                                                 32658    351001.583
/BNC/1.0/K/K5/K5D  &LSQB;UNCATALOGUED TEXT SAWLDA&RSQB;                  34572    352223.071
/BNC/1.0/H/HU/HU4  GUT$1$1$2JOURNAL OF GASTROENTEROLOGY AND HEPATOLOGY   29557    367084.662
/BNC/1.0/H/HW/HWS  GUT$1$1$2JOURNAL OF GASTROENTEROLOGY AND HEPATOLOGY   29531    367801.892
/BNC/1.0/H/HU/HU2  GUT$1$1$2JOURNAL OF GASTROENTEROLOGY AND HEPATOLOGY   29356    379109.473
/BNC/1.0/H/HU/HU3  GUT$1$1$2JOURNAL OF GASTROENTEROLOGY AND HEPATOLOGY   29500    382579.770
/BNC/1.0/H/HH/HHX  HANSARD PROCEEDINGS 199$1&NDASH;92 SESSION            30854    383471.711
/BNC/1.0/K/K9/K97  LIVERPOOL ECHO $1$21DAILY POST$1$1$2NOVEMBER 1992 --  38865    402754.606
/BNC/1.0/C/CR/CRM  NATURE                                                36867    463853.336
/BNC/1.0/H/HH/HHV  SELECTION FROM HANSARD 199$1&NDASH;1992               30389    469043.781
```

Each line shows the filename, the title of the text, the length of the contingency table and the value for $G^2$. These are sorted in ascending order since comparing two identical documents would produce a $G^2$ of zero. A brief inspection of the titles of the documents at the top of the list would indicate that the matching process has not been entirely successful. One way of evaluating this approach is to calculate the average rank of the 61 "Computergram International" texts, which are typical of the sort of texts this technique should identify. If the technique is working perfectly, the average rank should be 31. If it is completely random, the average rank would be 2062. It transpires that the average rank is in fact 1171.98, with std dev = 178.54. Clearly, this result is better than chance, but far from adequate. A number of modifications were therefore investigated.

The first modification was to eliminate those texts whose $G^2$ statistic was based on a very small contingency table. This can occur if the BNC text in question is particularly short (the BNC attempts to maintain a standard sample size but this was not always possible). A number of thresholds were investigated, and the optimum value (determined empirically) was around 23,560 words. When texts below this value are eliminated, the average rank of the Computergram International (CI) texts becomes 180.03, with std dev 81.27. It is possible to reduce this value still further, but only by compromising the overall recall value (i.e. genuine texts are eliminated along with the "noise").

The second modification was to reduce the contingency table to include just those words that appeared in both wfls. Since most of the 25,000 or so rows in a typical table contained mainly zeros in the second column, it was proposed that the $G^2$ calculation may be performed more effectively without them. However, when this modification was implemented, the overall pattern was almost identical to the previous trial; only the table size and $G^2$ value changed. Consequently, the average rank of the CI texts was very similar (mean = 1171.80, std dev = 175.18). However, as before, it is clear that the size of the contingency table is very small for some of the texts. So again a filter was applied to eliminate such "anomalies". The optimal value of the threshold was now 1370 (before recall was compromised), which produced an average rank of 125.15 (std. dev. = 75.62).

Although this evaluation has been quantitative, it is still somewhat anecdotal. To facilitate a more comparative evaluation, the bottom-up algorithm was re-applied to the BNC, using the rank correlation as the similarity measure. This produces the following top and bottom 10 texts:

| | | | |
|---|---|---|---|
| /BNC/1.0/H/HA/HAC | ARTICLES FROM PRACTICAL PC NOV 9$1&NDASH;FE | 7558 | 0.606 |
| /BNC/1.0/J/J0/J0V | ELECTRONIC INFORMATION RESOURCES AND THE HI | 5393 | 0.592 |
| /BNC/1.0/C/CT/CTX | WHAT PERSONAL COMPUTER | 4337 | 0.581 |
| /BNC/1.0/F/FT/FT8 | WHAT PERSONAL COMPUTER: THE ULTIMATE GUIDE | 4513 | 0.577 |
| /BNC/1.0/G/G0/G00 | MISCELLANEOUS ARTICLES ABOUT DESK-TOP PUBLI | 5228 | 0.572 |
| /BNC/1.0/C/CB/CBU | ACCOUNTANCY | 6053 | 0.567 |
| /BNC/1.0/C/CB/CBX | ACCOUNTANCY | 5902 | 0.557 |
| /BNC/1.0/K/KR/KRG | IDEAS IN ACTION PROGRAMMES (03) -- AN ELECT | 3231 | 0.556 |
| /BNC/1.0/H/HR/HRD | MULTIMEDIA IN THE 1990S | 3914 | 0.554 |
| /BNC/1.0/E/EE/EEB | PEOPLE IN ORGANISATIONS | 3199 | 0.553 |
| ..... | | | |
| /BNC/1.0/F/FU/FUS | RESULTS OF PRSTATECTOMY SURVEY -- AN ELECTR | 239 | 0.096 |
| /BNC/1.0/J/J5/J5B | ECOVER BIO-DEGRADABLE HOUSEHOLD CLEANING PR | 248 | 0.088 |
| /BNC/1.0/H/H4/H4P | MEDICAL CONSULTATIONS -- AN ELECTRONIC TRAN | 137 | 0.087 |
| /BNC/1.0/G/GY/GY5 | MEDICAL CONSULTATIONS -- AN ELECTRONIC TRAN | 90 | 0.078 |
| /BNC/1.0/K/KP/KPS | SPOKEN MATERIAL FROM RESPONDENT PAMELA2 -- | 21 | 0.063 |
| /BNC/1.0/G/G5/G54 | MEDICAL CONSULTATIONS -- AN ELECTRONIC TRAN | 17 | 0.052 |
| /BNC/1.0/H/HK/HKM | ROCKWELL/THE BETTER ALTERNATIVE TO THE FLAT | 215 | 0.050 |
| /BNC/1.0/F/FD/FDE | THE WEEKLY LAW REPORTS 1992 VOLUME 3&LSQB;P | 196 | 0.045 |
| /BNC/1.0/G/G5/G5A | AUCTION ROOMS -- AN ELECTRONIC TRANSCRIPTIO | 186 | 0.000 |
| /BNC/1.0/J/JJ/JJJ | BRISTOL UNIVERSITY -- AN ELECTRONIC TRANSCR | 18 | -0.259 |

Each line shows the filename, the title of the text, the number of common words and the value for r. Interestingly, although this is the more primitive technique, the results appear to be more intuitively satisfying. Of the top ten texts, six have titles that are clearly related to computing, including all of the top five. The remaining four could arguably be classified as Commerce & Finance (which was identified as the second most similar domain to email). However, a suitable title is no guarantee of suitable contents. As far as can reasonably be expected, the titles constitute a fair and accurate reflection of the contents of each text. Of course, the whole point of this approach is to develop reliable techniques that do not rely on ambiguous manual annotations such as title or domain, so the presence of suitable titles is merely an initial indication of success.

However, the quantitative analysis reveals a different picture. The value for the average rank of the CI texts is 959.852 (std dev = 524.435). Even when a threshold of 1370 is applied the average rank remains as high as 818.410 (std dev = 407.806). So despite the presence of a number of suitable candidates in the top 10, the overall performance of this technique (measured by the average rank of the 61 CI texts) is far weaker than that of the $G^2$ (which produced an average rank of 125.148 (std. dev. = 75.621). The $G^2$ statistic is therefore more suitable for this type of data since it uses the actual frequency values for the words in the wfls, rather than just their ranks. Furthermore, it is mathematically more well-grounded and produces results that appear to correspond reasonably well with human judgements of distinctiveness (Daille, 1995).

However, the rank correlation and log-likelihood statistics both make use of only unigram information. It is clear that much of the information that humans use to measure textual similarity is found not (solely) in the individual word frequencies (unigrams), but rather in the way they combine (n-grams). The logical next step is therefore to compare word bigrams (or trigrams) instead of just unigram data. A variation on this would be to compare texts using the log-likelihood applied to bigrams which are not necessarily adjacent, i.e. counting occurrences of word1 and word2 within a limiting distance of each other. Ironically, such methods have been previously used

for actually building the LMs themselves, and have been successfully applied to both speech (Rose & Lee, 1994) and handwriting data (Rose & Evett, 1995). Counting words within a limited window would be smoother than using strict bigrams and consequently less affected by the problems caused by sparse data (which are inevitable when small, individual text files are compared).

Another intriguing possibility is to use the LM itself as the similarity metric. After all, a LM can be applied to a test text to produce a perplexity score, which effectively measures how well the LM predicts the words in the text. So if the LM was trained from a text that is very similar to the test text, then it should predict the test data well and the perplexity should be low. Conversely, if the test text is very different from the training text, then the perplexity will be high. The perplexity score therefore can be used to measure textual similarity. Moreover, it has the advantage doing so by considering unigram, bigram and trigram data.

However, there are problems with this method. Firstly, the LM is being used as the representation of the training text against which similarity is to be judged, and yet it is, by definition, under-trained and therefore degenerate. Secondly, the method by which similarity is measured should be different from the method by which success is evaluated. Since the aim is to improve language model quality, the extent to which this is successful will be judged is by its perplexity. To use perplexity both as an improvement metric and an evaluation metric implies a certain amount of circular reasoning. However, the use of such iterative techniques is not totally without precedent within the LM community. Several research groups have reported the successful improvement of LMs using techniques that iteratively tune the LM parameters using new samples of training data (e.g. Jelinek, 1990). So, this approach may transpire to be sufficiently well principled to merit further investigation.

# *4. Language model quality*

A LM is built by collecting trigram, bigram & unigram data from a training corpus. However, it is not always desirable to store all of this data. Thresholds can be set such that some of the lower frequency n-grams are discarded. For example, a trigram cut-off of 5 implies that all the trigrams with frequencies of 5 or fewer in the training data are not used in building the model. Setting lower thresholds allows the model to focus on more frequent events, and produces a proportionately smaller model. The LMs described in this paper were built using the CMU SLM toolkit (Rosenfeld, 1994) which facilitated the construction of a variety of LMs representing a range of different settings for each of the pertinent parameters.

The first of these was the Email LM. This was constructed using a vocabulary of 20,000 words which was derived directly from the email training data. The bigram and trigram cutoffs were both set to zero. The second LM was built from the whole of the BNC, using the same vocabulary as the Email LM (in order to ensure consistency). In other words, although their n-grams had been based on general English rather than Email, their vocabulary was derived from the Email data. For comparison therefore, a third BNC LM was built, using a vocabulary derived directly from the BNC (rather than email). This allowed the comparative evaluation of the contribution of vocabulary vs. n-grams to the LM effectiveness (measured using both perplexity and word error rate). Due to memory constraints it was not possible to build the BNC models with cut-offs lower than 2-2. The fourth LM investigated was the 20k WSJ LM that is available from the Abbot ftp site at Cambridge University.

## 4.1 Evaluating the language models

Once a LM has been built, it is desirable to have some measure of its quality. One such measure is known as perplexity (PP), which can be thought of as a measure of the "branching factor" (i.e. the average size of the set of words between which the recogniser must choose) when transcribing a single word of the spoken text. PP thus measures the recognition difficulty of the text relative to the given LM, and is measured by applying the model to a sample of test data. This process therefore can be performed off-line, i.e. independently of the speech recogniser for which the models are intended. The testing was performed using the CMU toolkit, by applying each LM to the test data which in this case was a sample of 10,000 words from the transcriptions of a database of video mail messages, developed by Cambridge University as part of their Video Mail Retrieval using Voice project (Jones et al, 1994). Unfortunately, it was not possible to calculate the PP of the WSJ LM, due to the absence of a readily available version in the correct format.

A second evaluation method is to integrate the LM with the speech recogniser and test the combined system using recorded speech data. The models can be interchanged between trials, allowing comparative evaluation by measuring the word error rate (WER) produced by each model. More precisely, the error rates are measured using two standard metrics, percentage correct and accuracy:

1. $$\%Correct = \frac{H}{N} \times 100\%$$

2. $$Accuracy = \frac{(H - I)}{N} \times 100\%$$

where: H is the number of correct transcriptions (words in the utterance that are found in the transcription), D is the number of deletions (words in the utterance that are missing from the transcription), S is the number of substitutions (words in the utterance that are replaced by an incorrect word in the transcription), and I is the number of insertions (extra words in the transcription). Accuracy is more critical than %Correct in that it directly penalises insertions. Deletions & substitutions reduce the value of H, since H = N - (D+S).

As mentioned above, the VMR database is a collection of speech data with transcriptions (of which the latter were used in the above evaluation). The speech part contains audio files for 15 speakers, of which 10 were used in the current investigation. The Abbot recogniser was run using each combination of the 10 speakers' data files (as input) and each of the four LMs: email, BNC with email vocabulary, BNC and the WSJ LM. The output transcriptions were assessed for %correct and accuracy using the HResults program, which is part of HTK - the Hidden Markov Model toolkit (Young & Woodland, 1993).

| | %Correct | Accuracy | | %Correct | Accuracy |
|---|---|---|---|---|---|
| **Speaker 1** | | | **Speaker 6** | | |
| email | **42.32** | **33.23** | email | **52.17** | **42.88** |
| BNC | **41.47** | **32.02** | BNC/email | **49.53** | **39.76** |
| BNC/email | **41.42** | **31.06** | BNC | **48.97** | **39.11** |
| WSJ | **37.84** | **28.50** | WSJ | **46.36** | **36.65** |
| **Speaker 2** | | | **Speaker 7** | | |
| BNC/email | **37.14** | **24.58** | email | **65.05** | **54.23** |
| email | **36.77** | **25.03** | BNC/email | **64.49** | **53.93** |
| BNC | **36.19** | **24.40** | BNC | **64.23** | **53.83** |
| WSJ | **32.90** | **22.11** | WSJ | **60.97** | **50.15** |
| **Speaker 3** | | | **Speaker 8** | | |
| email | **44.42** | **39.14** | email | **54.70** | **43.44** |
| BNC/email | **44.42** | **37.86** | BNC/email | **51.40** | **38.74** |
| BNC | **43.34** | **36.74** | BNC | **50.99** | **37.64** |
| WSJ | **39.72** | **33.90** | WSJ | **47.93** | **35.55** |
| **Speaker 4** | | | **Speaker 9** | | |
| email | **59.82** | **50.04** | email | **70.08** | **62.75** |
| BNC/email | **59.28** | **49.50** | BNC/email | **69.86** | **62.03** |
| BNC | **58.37** | **48.28** | BNC | **68.85** | **61.14** |
| WSJ | **56.99** | **46.68** | WSJ | **66.67** | **58.37** |
| **Speaker 5** | | | **Speaker 10** | | |
| BNC | **76.52** | **71.99** | email | **65.91** | **56.14** |
| BNC/email | **75.94** | **70.84** | BNC/email | **65.83** | **55.38** |
| WSJ | **73.15** | **68.43** | BNC | **65.12** | **54.67** |
| email | **71.90** | **66.31** | WSJ | **61.58** | **51.13** |

| **OVERALL** | **%Correct** | **Accuracy** | **Perplexity** |
|---|---|---|---|
| email | **54.92** | **45.85** | **261.58** |
| BNC/email | **54.04** | **44.24** | **241.70** |
| BNC | **53.40** | **43.71** | **227.54** |
| WSJ | **50.42** | **40.93** | **N/A** |

**Table 2. %Correct, accuracy and perplexity of the language models**

Table 2 shows the results of this investigation. The results for %correct and accuracy show the combined effect of the recogniser and LM. The contribution of the LM depends on its vocabulary and perplexity. As the LM changes, it produces different behaviour in the combined system and therefore different types of errors (e.g. insertions, deletions & substitutions). The net effect is that the email LM produces the highest %correct and also the highest accuracy. It is around 5% better (on both measures) than the WSJ LM. This is remarkable, considering the tiny corpus from which it was derived (2 million vs. 227 million in the case of WSJ). In between these two extremes are the two BNC LMs - the one with the email vocabulary performs slightly better (~0.5%) than the one with its own vocabulary.

The result for the PP testing is highly interesting, although somewhat puzzling. As described earlier, a corpus of low homogeneity should produce a LM of higher PP than a corpus of high homogeneity. This is indeed shown to be the case, since the PP for email is 261.58 (homogeneity = 0.362), whereas the PP for the BNC is 227.54 (homogeneity = 0.687). These PP values are calculated using the 10K test data sample from the transcriptions of the VMR project. The higher PP value for email would tend to indicate that this is the poorer LM. However, it is clear that when these LMs are used on the real spoken data, the email LM provides the lowest error rates. Initial explanations for this centred on the vocabulary, since a higher incidence of out-of-vocabulary (OOV) words can produce a lower PP but a higher WER. However, the email LM performs better (by 0.88% correct) than the BNC/email LM even though both share the same vocabulary. There must therefore be n-grams in the email corpus that are simply not found in the BNC (even though the BNC is 50 times larger). This implies that quality, not quantity, is a major factor in training effective LMs. Further PP testing, possibly using the complete transcriptions of the VMR data is necessary to clarify this issue.

Evidently, the choice of vocabulary also makes an important contribution. The BNC LM with the email vocabulary performs better (by 0.64% correct) than the BNC LM with its own vocabulary, so clearly the email vocabulary provides better coverage of the test data. In fact, it is possible to directly compare the OOV rates with the performances shown above: the BNC LM with the email vocabulary has an OOV rate of 1.16% on the VMR data, and a %correct of 54.04. By contrast, the BNC LM with its own vocabulary has an OOV rate of 1.69% and a %correct of 53.40. These figures suggest that an increase in OOV rate of 0.56% leads to a reduction in %correct of 0.64%, or, in other words, a 1% increase in OOV rate produces a reduction in %correct of around 1.14%. Interestingly, this figure correlates extremely well with the results of a similar experiment performed by Rosenfeld (1995), who found that a 1% increase in the OOV rate can lead to a 1.2% increase in the word error rate.

## 5. Conclusions

The analysis of the corpora has provided several revealing insights. Firstly, it is necessary to determine the homogeneity of a corpus prior to performing any similarity measures, since it is not clear what a measure of similarity would mean if a homogeneous corpus was being compared with a heterogeneous one. Clearly, the email corpus is highly heterogeneous. This means it is particularly prone to "burstiness" and unpredictability, which affects all levels of n-grams (including unigrams). This may be due in part to the particular training corpus used, but it is more likely to be something inherent to the medium, since email can fulfil so many communicative functions. It therefore exhibits a level of diversity surpassed perhaps only by spontaneous speech. Investigation of the spoken part of the BNC is therefore suggested as an area for further work.

To a certain extent the apparent heterogeneity of the email undermines the results of the similarity measure, which shows email to be very different to the BNC. Nevertheless, the extent to which the email is unlike all the other BNC domains is quite apparent and therefore mitigates any unprincipled approaches to corpus augmentation using crude, top-down techniques that involve complete domains taken from the BNC. Consequently, the best way to acquire more email data appears to be either: (a) instigate a further collection initiative, or (b) use more sophisticated bottom-up methods. The similarity metric used in (b) must be chosen carefully. Although the log-likelihood and rank correlation metrics both produce results that can look intuitively plausible, this merely underlines the need for an objective, thorough evaluation method. Log-likelihood is the more principled of the two measures, and it is likely that this offers the greater potential.

The results of the language modelling exercise provide clear evidence that it is possible to build effective LMs from small corpora. The email LM outperformed the other LMs on real spoken data (albeit taken from a technical, "email-like" domain) for eight of the ten speakers. This is remarkable, considering the other LMs are trained on corpora that are several times larger. This effect can be partially attributed to the source of the n-grams: the email LM performs better than the BNC/email LM even though both share the same vocabulary, so there must be n-grams in the email corpus that are simply not found in the BNC. However, the choice of vocabulary is also important: the BNC/email LM performs better than the BNC LM, so clearly the email vocabulary provides better coverage of the test data.

### 5.1 Further work

It is possible to personalise LMs dynamically using cache-based methods (e.g. Kuhn & de Mori 1990) and the evidence suggests that this may prove the more effective approach (Matsunaga et al, 1992). To develop an optimum LM, the long term strategy must be to develop a dynamic LM. It is clear the email is highly heterogeneous, and therefore inherently unpredictable. Attempting to model this by static means can therefore produce only limited success. A dynamic LM would adapt to the current input, and update its probabilities accordingly. The starting point for a dynamic LM could be a static model based on a combination of the email LM and the BNC LM. Such an interpolated model could, with appropriately derived interpolation weights, provide the best starting point for the dynamic LM.

There are still a number of open issues regarding the choice of vocabulary. Clearly this is a crucial parameter, and building a vocabulary simply by taking the top N words of the BNC is crude and sub-optimal. It is suggested that a 1% increase in the OOV rate can lead to a 1.2% increase in the word error rate (Rosenfeld, 1995). It is therefore necessary to investigate more sophisticated vocabulary selection techniques (e.g. Jelinek 1990). However, the process of bottom-up corpus augmentation may go some way to alleviating these problems, since an appropriately extended training corpus should provide more reliable unigram frequency information.

## *6. References*


Biber, D. (1993) "Using register-diversified corpora for general language studies", Computational Linguistics, 19, No. 2

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., P. (1989) "A statistical approach to machine translation", Technical Report, IBM Research Division

Burnard, L. (1995) Users Reference Guide for the British National Corpus, Oxford University Computing Services

Church, K., Hanks, P., Hindle, D. and Gale, W. (1991) "Using statistics in lexical analysis", in Zernik, U. (Ed.) "Lexical Acquisition: Using On-Line Resources to Build a Lexicon", LEA, NJ.

Daille, B. (1995) "Combined approach for terminology extraction", Technical Report 5, UCREL, Lancaster University

Dunning, E. (1993) "Accurate methods for the statistics of surprise and coincidence", Computational Linguistics, 19, No. 1

Hochberg, M., Robinson T. & Renals S. (1994) "Large vocabulary continuous speech recognition using a hybrid connectionist HMM system", Proc. of ICSLP, pp. 1499-1502

Jelinek, F. (1990) "Self-organized language modeling for speech recognition", in Waibel and Lee (Eds.), Readings in Speech Recognition, Morgan Kaufmann, San Mateo, CA

Jones, G., Foote, J., Sparck Jones, K. & Young, S. (1994) "Video mail retrieval using voice", Technical Report 335, Cambridge University Computer Laboratory

Kilgarriff, A. (1996) "Which words are particularly characteristic of a text?" in Evett & Rose (Eds.) Language Engineering for Document Analysis & Recognition, AISB Workshop proceedings

Kuhn, R. & De Mori, R. (1990) "A cache-based natural language model for speech recognition" IEEE Trans. on PAMI, 12(6), pp. 570-583

Lewis, D. (1992) "Text representation for intelligent text retrieval: a classification-oriented view", in P. Jacobs "Text-based Intelligent Systems", LEA Publishers, Hillsdale, NJ.

Matsunaga, S., Yamada, T. & Shikano, K. (1992) "Task adaptation in stochastic language models for continuous speech recognition", Proc. ICASSP Vol. 1, pp. 165-168

Rose, T.G. & Evett, L. (1995) "The use of context in cursive script recognition", Machine Vision and Applications", Springer International

Rose, T.G. & Lee, M (1994) "Language modelling for large vocabulary speech recognition", Proc. IOA Meeting on LVSR, Cambridge, England

Rosenfeld, R. (1994) "The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARPA CSR Evaluation, Proceedings of the Spoken Language Technology Workshop 1995, Austin (TX)

Rosenfeld, R. (1995) "Optimizing lexical and n-gram coverage via judicious use of linguistic data", Proc. Eurospeech 95

Rudnicky, A. (1995) "Language modeling with limited domain data", Proceedings of the ARPA Workshop on Spoken Language Technology, Morgan Kaufmann, San Mateo, pp. 66-69

Vergyri, D. (1995) Unpublished web page http://www.clsp.jhu.edu/~dverg/bleaching.html

Young S. & Woodland P. (1993) "HTK: Hidden Markov Model Toolkit V1.5 User Manual", Cambridge University Engineering Dept. and Entropic Research Labs Ltd.