ON THE FEASIBILITY OF USING APPLICATION-SPECIFIC SPEECH TO DERIVE A GENERAL-PURPOSE SPEECH RECOGNISER TRAINING DATABASE

M O'Kane and P Kenne
Faculty of Information Sciences and Engineering
University of Canberra

ABSTRACT - A speech database is easier to mark-up if what the speakers are saying is known before marking-up commences. In a joint project with the court reporting services we have examined the feasibility of using court speech recordings and associated transcript to derive a general-purpose speech recogniser training database. The first question addressed was the size of the natural vocabulary that was covered by day-to-day court proceedings. The next question addressed was the frequency of occurrence of the various words and phrases in this vocabulary. We then turned to the issue of how much transcript had to be examined in total in order to get a reasonable number of examples of all the commonly-occurring words in the vocabulary. All this work was done using automatic analysis of transcript text.

Another important aspect of speech-database collecting is the overall time it is going to take to mark-up a database of known size. In order to address this issue we conducted mark-up speed trials in which several experienced speech database markers were timed for speed of marking-up speech from associated transcript. A special software mark-up system was used which ideally requires only four mouse-clicks to mark up and confirm each instance of each word entered in the database. Each marker was marking-up at word level only. Quality of marking-up was checked for each marker.

While the exact minimum amount of data needed to train a very large speech recogniser is unknown, experiments such as the ones described here suggest that the concept of deriving such databases from application-specific speech is a very large but not an impossible task.

INTRODUCTION

Practical use of statistically-based recognisers requires the generation of large marked-up databases to train the recognisers. The collection and marking-up of speech databases is a time-consuming task. Investigating ways of speeding up the collection of speech-recogniser training databases, we examined court transcript data. Court and parliamentary transcript has been used by others for language-modelling studies (Brown, della Pietra, Mercer, della Pietra and Lai, 1992). In Australian courts essentially all proceedings are recorded and the recordings are retained for a statutory period. For most courts written transcript is also produced and retained. We investigated the issue of seeing if the combination of transcript and recording could be used to derive a useable database for training general-purpose and application-specific speech recognisers.

In order to gain some idea of the vocabulary size and the number of repetitions of items in the vocabulary we considered eight consecutive days of transcript from one on-going case in one court. The growth in vocabulary size as a function of the number of days of transcript analysed is given in Figure 1a. In Figure 1b this information is presented again on the same scale as the graph of the total number of all words processed as a function of the number of days of transcript analysed. (It should be noted that the transcript size for different days varies.) It can be seen that the growth in the vocabulary is quite small. In total after processing eight days of transcript (170,638 words) the vocabulary is 6133. If all words were repeated equally often this would mean that the average repetitions per word was approximately 28.

Of course all words do not occur equally often. So we next investigated the high-frequency words and the number of times they were repeated. The ten highest-frequency words and their number of occurrences are given in Table 1. Note that the word "the" accounts for 6% of all the words processed and the words ranked 2-10 in the top ten occurring word list together account for another 22.5% of all words processed.

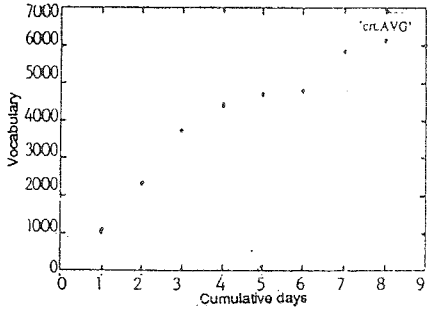

Figure 1a: Growth in vocabulary as a function of the number of days of transcript

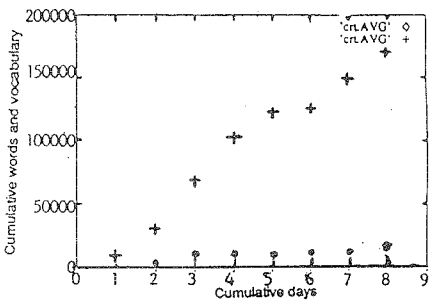

Figure 1b: Number of words processed as a function of number of days of transcript. Vocabulary growth also shown for comparison

word	occurrences
the	10,260
that	5,041
to	4,874
of	4,687
and	4,645
I	4,004
you	3,841
in	3,670
is	2,940
it	2,828

Table 1: The ten most-frequently occurring words in 8 days of transcript

word	occurrences
of the	1,400
in the	994
to the	835
your honour	563
do you	558
it is	502
i think	460
is that	450
and the	414
yes and	410

Table 2: The ten most-frequently occurring word pairs in 8 days of transcript

We were interested not only in the number of items in the vocabulary but also the number of word pairs, word triples, word quadruples and word quintuples that could occur in the transcript and how these numbers varied as each new day's transcript was added to the collection. This information is presented in Figure 2. For comparison with Table 1 the top ten occurring word pairs are given in Table 2.

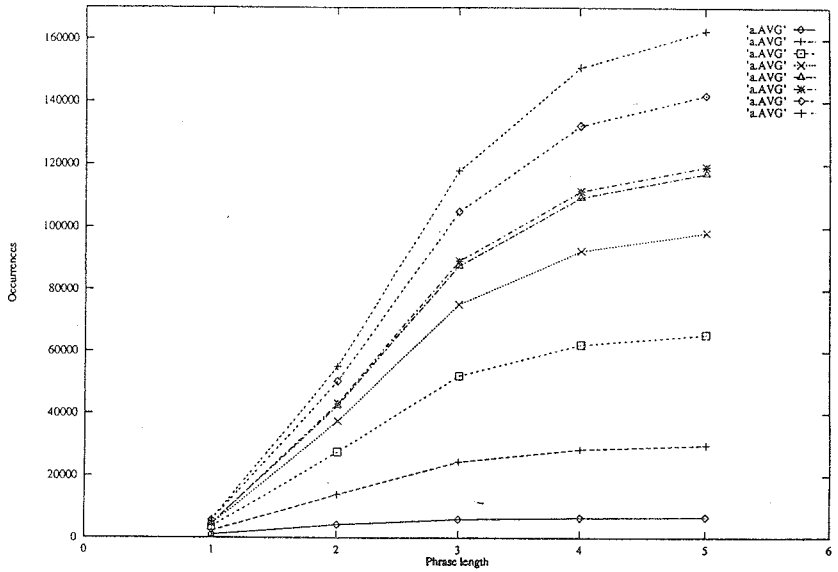

Figure 2: Number of different single words, word pairs, word triples, word quadruples and word quintriples as a function of number of days of transcript

The top ten words even though they do occur very often do not provide a very useful vocabulary, taken as a set. In order to see if there are enough repetitions of enough other words we considered, for the eight days of transcript, how many words had greater than 10000, 5000, 2000, 1000, 500, 200, 100 occurrences. This information is given in Table 3.

greater than 10000 occurrences	1
greater than 5000 occurrences	2
greater than 2000 occurrences	12
greater than 1000 occurrences	27
greater than 500 occurrences	54
greater than 200 occurrences	98
greater than 100 occurrences	229

Table 3: Number of words in 8 days of transcript which were repeated more than 10000, 5000, 2000, 1000, 500, 200 and 100 times

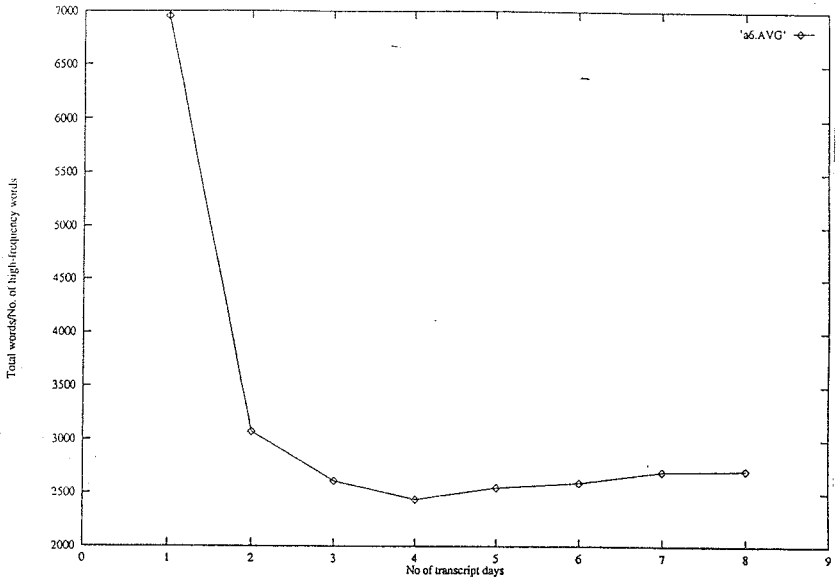

Figure 3: (Number of words processed)/(Number of words averaging >400 times) as a function of number of days of transcript

We were also interested in finding out how long it took to get a reasonable number of words of reasonably high repetition rate. Some idea of this is given in Figure 3 in which the total number of words processed divided by the number of words occurring more than 400 times is plotted as a function of number of transcript days. It is interesting that after day 4 (102,255 words processed), the ratio starts to rise, indicating that the high-frequency words are occurring more frequently but not a lot of other words are occurring more frequently.

GENERAL VERSUS SPECIFIC

One would expect that collecting a lot of data from one jurisdiction would lead to a large number of jurisdiction-specific words. Accordingly we investigated for the top 10, 20, 30, 40 and 50 occurrences of the vocabulary, the word pairs, the word triples, the word quadruples and the word quintuples what the ratio of general words or phrases was to the total number of words or phrases for each set. This information is given graphically in Figure 4. The single words are almost totally general at this part of the file with the number of jurisdiction-specific phrases becoming more frequent as the phrase gets longer.

SPEAKER VARIETY

The number of speakers contributing to any given transcript can vary enormously. For the eight days considered in the experiment described above, only five speakers (a judge and four counsel) contributed to the first seven days while fourteen speakers contributed to the eighth day which was the day on which witnesses were called for the first time.

In order to see if data from a single court was comparable to data from another source we analysed two days of conference proceedings and compared data from these with data from the two longest days (day 3 and day 4) of the eight day sequence described above. This information is given in Table 4.

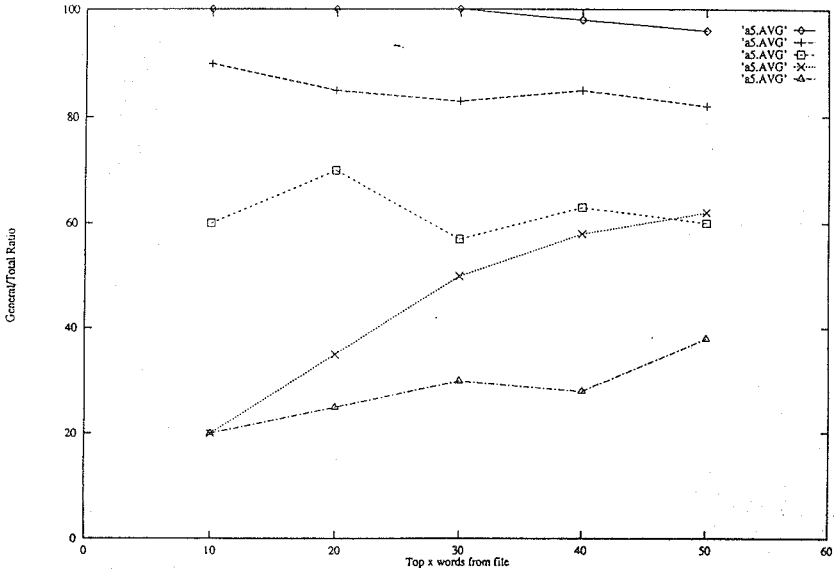

Figure 4: Ratio of general/total for single words, word pairs, word triples, word quadruples or word quintriples for the top 10,20,30,40 and 50 most-frequent occurrences

	Conference Day 1	Conference Day 2	Court Day 3	Court Day 4
Total number of words in transcript	53478	37629	37043	34520
No. of words in vocabulary	4191	3563	2719	2462
No. word pairs	26221	19788	17246	15847
No. word triples	44988	32195	29788	27477
No. word quadruples	51347	36283	34438	31859
No. word quintuplets	52883	37304	35968	33359

Table 4: Comparison of statistics for conference and court days

MARK-UP TIME

The issue that makes this series of experiments worthwhile is that a court-derived speech database is extremely easy and quick to mark-up using a special software tool in which the speech appears in one window and the associated transcript in another. To mark-up a word, the operator uses only four mouse clicks - one to mark the start of the word, one to mark the end, one to press to play the word and thus aurally confirm the start and end, and one in the transcript window to save the marks.

Mark-up trials by various markers on long speech files indicate that a skilled and accurate marker averages about 14 seconds to mark-up a word in this way. This allows for fixing errors and scrolling both windows when necessary.

HOW FEASIBLE?

Although the returns are considerable, the marking-up of court transcript is still a daunting task. For example it would take an operator 16½ weeks to mark-up the eight days of transcript, assuming the marking-up rate of 14 seconds per word and a 40-hour working week.

Where one starts to win however is by a bootstrapping technique, whereby initially all words are marked but when enough repetitions (whatever 'enough' means) of any word or phrase have been marked these are used to train a statistical wordspotter which then marks-up these words and phrases automatically.

ACKNOWLEDGEMENT

This data analysed in this paper was provided by Auscript, the Commonwealth Reporting Service.

REFERENCE

Brown, P, della Pietra, V, Mercer, R L , della Pietra, S A and Lai J C (1992), *An estimate of an upper bound for the entropy of English,* Comp. Ling, 18, 31-40.