AN AUSTRALIAN SPEECH DATABASE DERIVED FROM COURT RECORDINGS

P.E.Kenne, M.J.O'Kane and H.Pearcy
The University of Adelaide

Abstract. We describe a pilot version and give a statistical characterisation of a large speech and natural language database which is being developed, and is based on the recordings of court proceedings and their transcripts.

INTRODUCTION

The value of large speech and natural language databases is now widely recognised within the speech processing and natural language (NL) communities, as can be seen by recent collection efforts, for example, MADCOW (Hirschman 1992), the OGI databases (Muthusamy, Cole and Oshika 1992), the Australian English database (Croot, Fletcher and Harrington 1992) and the Wall Street Journal corpus (Pane and Baker 1992). Text-and-speech aligned databases are useful as sources of naturally-produced training data for automatic speech recognisers. Databases of large amounts of spoken language also provide the raw data for a much deeper understanding of spoken language in its own right. In this paper a database currently being derived from court recordings and their transcripts and consider ways in which this corpus can be characterised statistically.

The origins of this project came from an enquiry from Auscript (the Commonwealth of Australia court reporting service) if it would be possible to use automatic speech recognition techniques to produce court transcripts. At present, all court proceedings are produced manually. A common technique is for shorthand reporters to take down the proceedings using shorthand, and then dictate their transcripts to typists. Partial automation has been introduced by having the shorthand reporters dictate their transcript to a speaker-dependent speech recognition system (e.g. such as produced by Dragon, Kurzweil and Verbex).

Courts are located in many different styles of rooms, and the acoustic conditions are highly variable. In addition, there has been no attempt to standardize the recording equipment until recently, when the decision was made to use DAT technology and to standardize the type of microphone used. Microphones are positioned in front of the judge, the lawyers for each side have a separate microphone, and there is a microphone located in the witness box. Each of these parties is recorded on a separate track. Even with high quality noise-cancelling microphones, the head and body movements of a lawyer, the judge or a witness results in variable quality recording.

Annually approximately 70000 hours of court proceedings are recorded in 250 courts (including various non-court organizations for which Auscript also reports), with approximately 800000 pages of transcript being produced. (Not all recordings are transcribed as a matter of course.) The court recordings, together with the transcripts, provide a rich source of data for speech and NL training, for example, it provides a source of natural emotive speech, which has been lacking in many speech databases (Eskénazi 1993). A difficulty in using this data to derive a speech recognition training database is that the transcripts are not in any way time aligned with the audio data. A further difficulty is that, due to Auscript editorial policy, the transcripts are not a faithful representation (neither at the word nor the phonetic level) of the audio.

STATISTICAL CHARACTERISATION OF THE COURT DATABASE

Some statistical and graphical characterisation of the court database has been given previously (O'Kane and Kenne, 1992). A typical court vocabulary is about 5000-20000 distinct words, and of these, about 350-600 words account for 80% of all words spoken in these cases. This effect is not peculiar to court speech alone. Figure 1 shows that similar effects hold for small-vocabulary telephone dialogues and for text examples such as novels. Relatively small numbers of word pairs, triples, quadruples and quintuples also account for a large proportion of coverage of the total transcript although this is not as pronounced as for single words (see Figure 2).

In any given case, there are a number of high frequency case specific words, and also a number of high frequency words specific to a particular stage of the court proceedings. For example, in the first day of a case involving a dispute between a number of unions workers, the words "application" and "crossapplication" occur frequently, but occur far less frequently on subsequent days, and phrases such

as "your honour" occur with high frequency over all days of the case). The types of speakers can be generally characterized as a small number of speakers who say a lot (the judge and lawyers), and a larger number of speakers who say relatively little (the witnesses). For example, in the case referred to above, the lawyers account for 64.5% of the utterances, the witnesses for 25.5% and the judge for 10%. The size of the lexicon varies according to the category of speaker. In this particular case, the judge has a vocabulary size of 1679 words, the lawyers together have a vocabulary of 4475 words, and the witnesses have a vocabulary of 3203 words. Similar observations may be made about other cases in the database. Tables 1 and 2 below characterize two cases in the database.

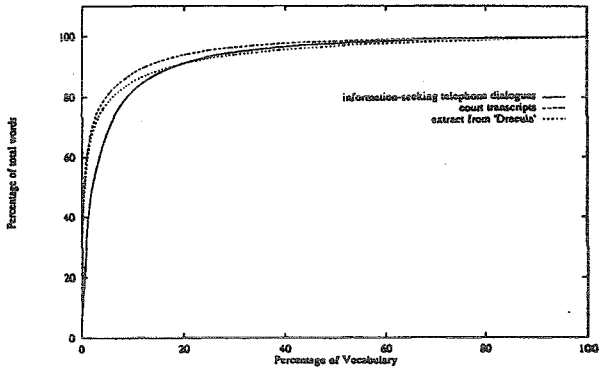

Figure 1: Total number of words in text accounted for as a function of percentage of total vocabulary represented by these words for three different corpora (Reproduced from O'Kane, 1993).

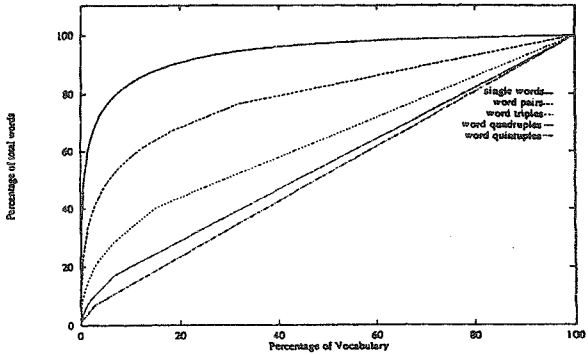

Figure 2: Total number of words, word pairs, triples, quadruples and quintuples in text accounted for as a function of percentage of distinct words, word pairs etc., these words, word pairs etc., represent for five days of court transcripts (Reproduced from O'Kane, 1993).

	Distinct Words	Total Words	Total Questions	Total Statements	Average utterance length
Judge	1679	17122	88	1027	15.3
Lawyers					
L1	1611	14699	614	122	19.9
L2	1432	10217	200	173	27.4
L3	2385	36805	1112	594	21.6
L4	3060	47921	1497	646	22.4
L5	325	859	0	11	78.0
L6	16	16	0	1	16.0
Witnesses					
W1	114	215	0	11	19.5
W2	53	76	0	4	19.0
W3	438	1729	0	169	10.2
W4	819	3798	0	219	17.3
W5	1004	5896	0	540	10.8
W6	729	5197	0	580	9.0
W7	216	550	0	51	10.8
W8	419	1861	0	168	11.0
W9	2346	24333	0	1821	13.4

Table 1: Data for case 1

	Distinct Words	Total Words	Total Questions	Total Statements	Average Utterance Length
Judge 2	1162	8369	48	606	12.8
Lawyers					
L7	1299	12924	362	142	25.6
L8	1931	18186	507	257	23.8
L9	4887	116920	4421	264	25.0
L10	3409	52544	2214	132	22.4
Witnesses					
W10	55	383	0	83	4.6
W11	1448	9366	0	555	16.9
W12	376	1457	0	124	11.8
W13	591	2290	0	245	9.4
w14	361	1152	0	95	12.1
w15	682	3352	0	130	25.8
w16	1405	9395	0	682	13.8
w17	831	3906	0	266	14.7
w18	641	2563	0	109	23.5
w19	3489	55851	0	3002	18.6
w20	276	887	0	71	12.5
w21	171	380	0	71	5.4
w22	235	549	0	44	12.5
w23	486	1901	0	153	12.4
w24	785	4733	0	339	14.0
w25	214	467	0	45	10.4
w26	1119	6502	0	540	12.0
w27	1449	15447	0	792	19.5

Table 2: Data for case 2

The tables below provide summary information for nine cases from the database. Table 3 gives the number of words, the number of utterances and the average utterance length (in words) for each of the cases, was well as the number of short (less than five words) utterances. Table 4 gives the number of short utterances as a percentage of all utterances by case. Table 5 shows the number of distinct one, two and three word utterances by case.

Case	Total words	Total utterances	Average utterance length	1 word utterances	2 word utterances	3 word utterances	4 word utterances
c1	171265	9831	17.42	1101	505	659	506
c2	331915	16720	19.85	1730	762	906	660
c3	41738	2764	15.10	209	137	188	152
c4	76136	5805	13.12	754	360	432	335
c5	152516	2733	55.81	137	72	102	60
c6	115486	5221	22.12	699	180	232	224
c7	116679	6471	18.03	1020	341	449	324
ca	674438	32780	20.57	3892	1565	2065	1615
c8	1707456	66954	25.50	7696	3225	3063	2656

Table 3: Summary data for all cases

Utterance length	c1	c2	c3	c4	c5	c6	c7	ca	c8
1	11.20	10.35	7.56	12.99	5.01	13.38	15.76	11.87	11.49
2	5.14	4.56	4.97	6.20	2.63	3.45	5.27	4.77	4.82
3	6.70	5.42	6.80	7.44	3.73	4.44	6.94	6.30	4.57
4	5.15	3.95	5.50	5.77	2.20	4.29	5.01	4.93	3.97

Table 4: Short utterances as a percentage of all utterances by case

Utterance length	c1	c2	c3	c4	c5	c6	c7	ca	c8
1	82	99	24	72	19	47	59	224	336
2	226	253	70	157	35	98	164	641	1054
3	388	460	119	272	46	160	224	1483	

Table 5: Number of distinct short utterances by case

Tables 6 and 7 show the most frequent one and two word utterances by case. Table 8 gives the number and percentage of utterances which start with the listed words.

	c1	c2	c3	c4	c5	c6	c7	ca	c8
yes	801	1247	148	456	99	475	804	2765	5532
no	106	175	20	110	11	65	89	398	364
mm	30	70	14	24	-	11	14	93	91
right	8	38	2	46	-	23	3	82	359
well	18	24	2	11	5	13	25	71	171
sorry	8	8	3	9	2	15	2	39	86
okay	1	6	1	2	2	26	2	34	81
correct	23	24	-	1	-	3	6	33	327
sure	-	3	-	-	-	15	3	18	32
certainly	3	8	-	9	3	5	-	18	30

Table 6: Top one word utterances by case

	c1	c2	c3	c4	c5	c6	c7	ca	c8
that's_correct	59	198	39	14	-	5	21	138	396
that's_right	21	98	-	40	-	16	61	138	120
thank_you	33	198	7	56	12	8	9	110	217
yes_yes	21	34	6	9	3	8	43	89	128
all_right	4	49	1	11	-	36	9	61	420
i_see	24	3	2	10	7	2	11	56	59
no_no	14	5	1	5	1	3	1	25	26
oh_yes	4	6	1	1	-	-	13	19	54
i_do	9	9	1	6	-	2	1	19	32
i_am	12	2	-	1	-	1	1	15	22
it_is	11	7	-	1	-	1	-	13	22

Table 7: Top two word utterances by case

	c1	c2	c3	c4	c5	c6	c7	ca	c8
yes	18.91	15.05	16.28	18.48	12.07	17.56	23.55	18.64	17.64
well	9.15	6.58	10.75	5.98	4.54	6.49	6.60	7.34	7.78
I	9.73	8.48	11.47	7.58	5.31	7.39	6.83	8.11	7.05
and	7.91	10.41	7.34	8.80	8.74	7.72	7.08	7.90	4.42
no	5.17	6.20	3.91	6.36	2.38	5.86	4.44	3.98	3.56
but	0.99	2.19	0.76	1.17	1.02	1.61	2.16	1.33	2.00
You	1.69	2.92	1.30	1.72	0.91	2.60	2.06	1.81	1.57
I'm	0.51	0.59	0.83	1.07	8.78	0.38	2.47	0.58	1.17
do(es)	1.87	0.82	2.10	1.03	0.22	1.11	0.70	1.25	1.04
what('s)	1.17	0.78	1.77	1.26	0.99	0.92	1.10	1.20	0.89
is(n't)	0.81	0.89	0.65	0.59	0.59	0.73	0.60	0.68	0.78
correct	0.24	0.17	0	0.02	0	0.10	0.09	0.11	0.58
They	0.68	0.50	0.43	0.40	0.26	0.33	0.46	0.47	0.46
which	0.40	0.18	0.22	0.09	0.29	0.42	0.34	0.31	0.33
can	0.51	0.77	0.62	0.31	0.51	0.86	0.15	0.47	0.31
how	0.46	0.15	0.43	0.78	0.11	0.23	0.40	0.44	0.26
I've	0.21	0.16	0.22	0.19	0.18	0.15	2.01	0.20	0.23
when	0.59	0.34	0.76	0.65	0.37	0.36	0.42	0.52	0.21
You're	0.08	0.04	0.14	0.14	0.22	0.04	0	0.09	0.17
why	0.14	0.22	0.14	0.17	0.11	0.11	0.09	0.13	0.17
mm(m)	0.38	0.44	0.54	0.53	0.04	0.21	0.23	0.34	0.16
You've	0.04	0.02	0.04	0.17	0.07	0	0	0.05	0.13
where('s)	0.28	0.09	0.11	0.24	0.18	0.06	0.12	0.19	0.13
They're	0.09	0.05	0.11	0.09	0.04	0	0	0.05	0.09
who('s)	0.36	0.08	0.47	0.41	0.11	0.08	0.22	0.28	0.05
They've	0.01	0	0.04	0	0	0.04	0	0.01	0.01

Table 8: Percentage of utterance starting with a given word

COMMON CHARACTERISTICS OF CASES

Considering the different court cases, it is interesting to note common features between them. For example if one considers the common most 20 words, then finds that the 20-word sets, these sets almost always have the same members and all occurrences of these words together account for a similar proportion of all word in each case (see Figure 3).

Similarly if one considers the utterance-initial words, the words 'yes', 'well', 'I', 'and', 'no' and 'but' account for more than 30% of all utterance-initial words. In the case of the single-word utterances, the words 'yes' and 'no' account for 6-14% of these utterances.

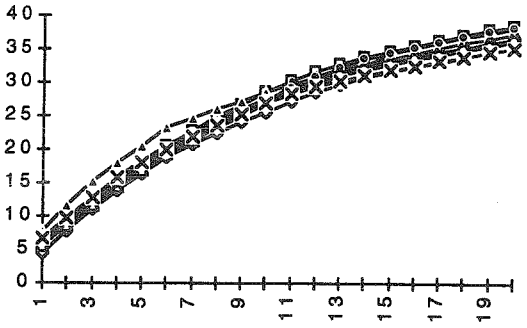

Figure 3: Percentage coverage of total words in cases by 20 most-frequently-occurring words in all nine cases.

REFERENCES

Croot, K., Fletcher, J. & Harrington, J. (1992), *Levels of segmentation and labelling in the Australian National Database of spoken language*, Proceedings SST 92, Brisbane, Australia, December 1992, 86-90.

Eskénazi, M. (1993), *Trends in speaking styles research*, Proceedings Eurospeech 93, Berlin, September 1993, 501-509.

Hirschman, L. (1992), *Multi-site data collection for a spoken language corpus MADCOW*, Proceedings ICSLP 92, Banff, Alberta, Canada, October 1992, 903-906.

Muthusamy, Y.K., Cole, R.A. & Oshika, B.T. (1992), *The OGI multi-language telephone speech corpus*, Proceedings ICSLP 92, Banff, October 1992, 895-898.

O'Kane M. J. (1993) "Listening intelligently and giving a sensible answer" in *AI'93* pp. 3-11 (World Scientific: Singapore; C. Rowles, H. Liu and N. Foo, eds.).

O'Kane M. & Kenne, P. (1993), *On the feasibility of using application-specific speech to derive a general-purpose speech recogniser training database*, Proceedings SST 92, Brisbane, Australia, December 1992, 712-717.

Pane, D.B. & Baker, J.M. (1992), *The design for the Wall Street Journal-based CSR corpus*, Proceedings ICSLP 92, Banff, Alberta, Canada, October 1992, 899-902.