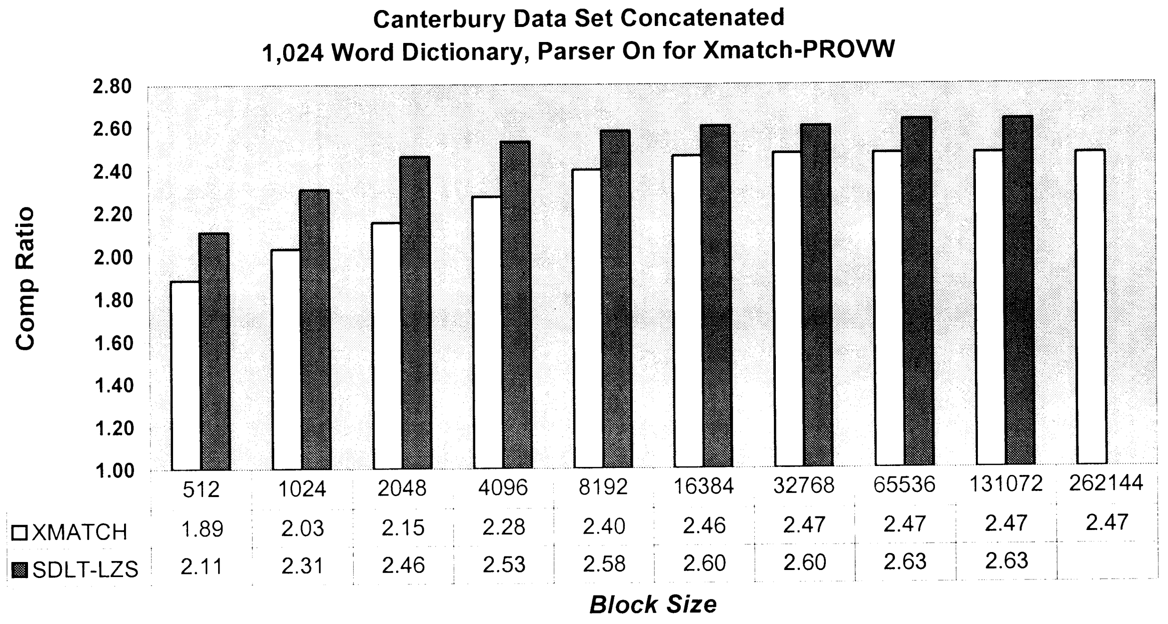
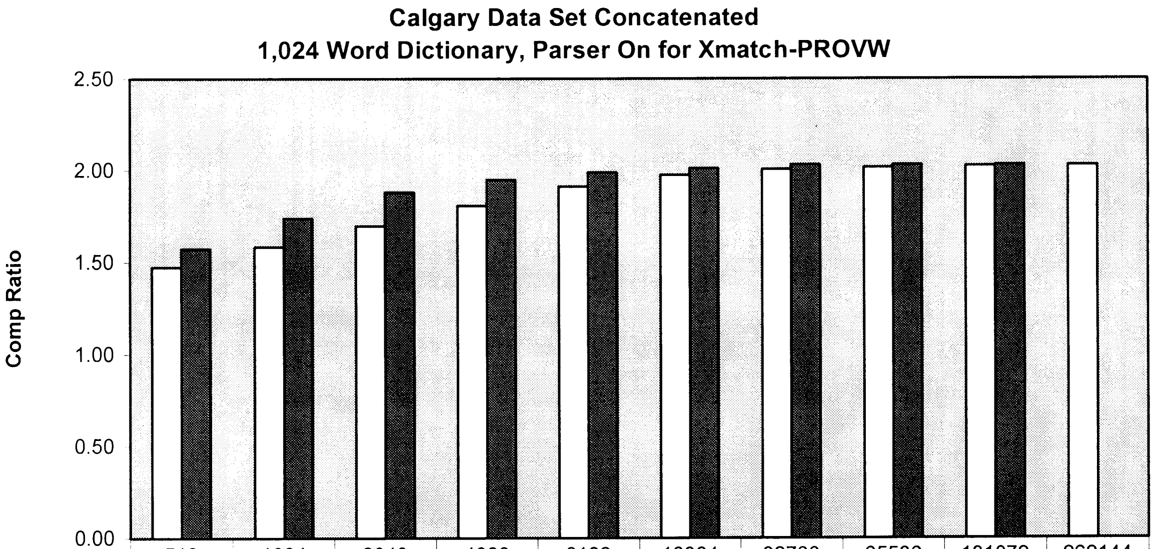
9.3. University of Canterbury Data Sets Comparison

The files in each of the data sets below have been concatenated into one file and their compression ratios measured.



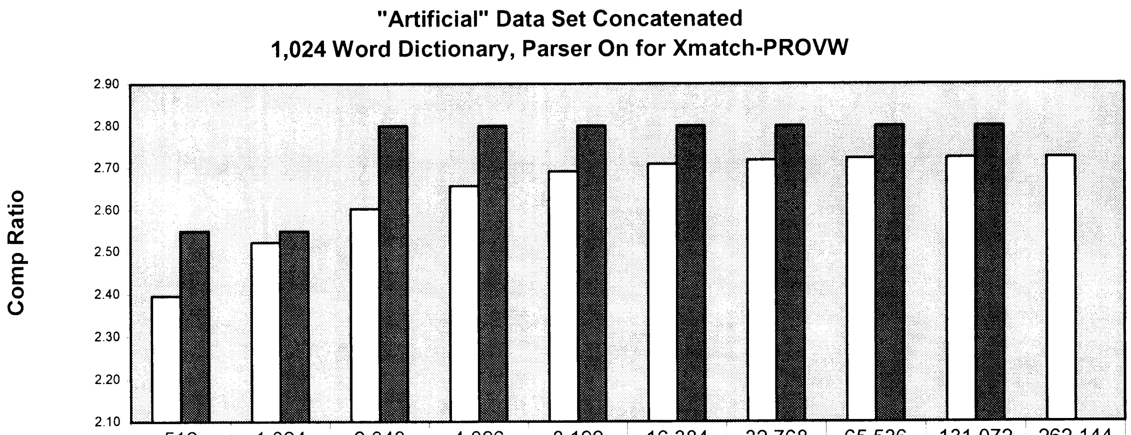
	512	1024	2048	4096	8192	16384	32768	65536	131072	262144
□XMATCH	1.89	2.03	2.15	2.28	2.40	2.46	2.47	2.47	2.47	2.47
■SDLT-LZS	2.11	2.31	2.46	2.53	2.58	2.60	2.60	2.63	2.63	

Table 25 - University of Canterbury - Canterbury Data Set Concatenated



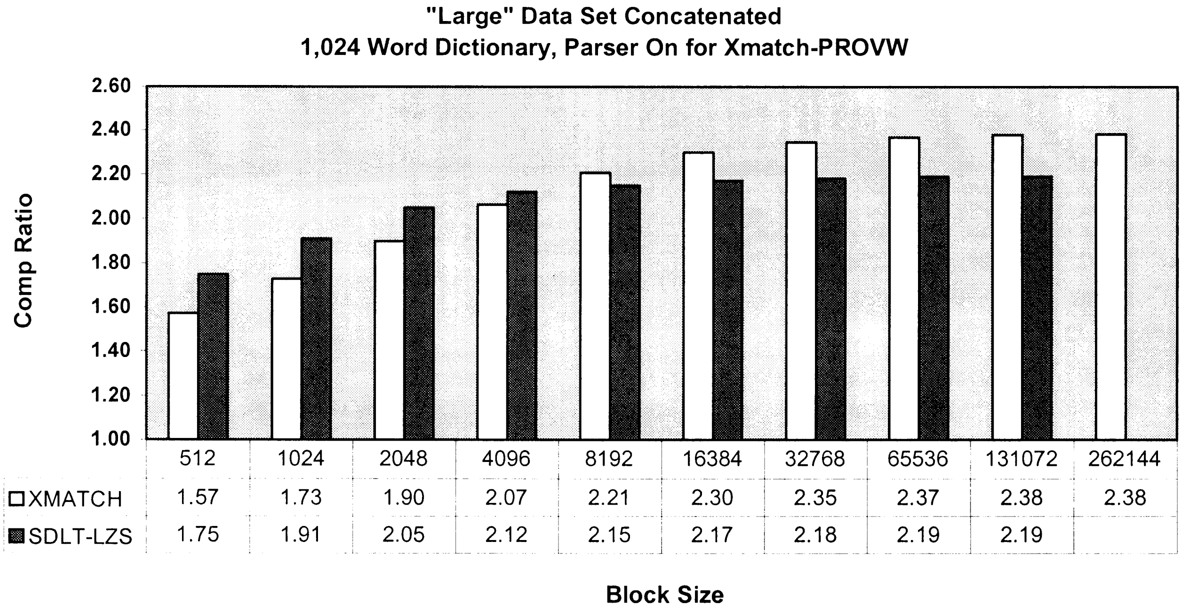
	512	1024	2048	4096	8192	16384	32768	65536	131072	262144
□XMATCH	1.47	1.58	1.70	1.81	1.91	1.97	2.00	2.02	2.02	2.03
▩SDLT-LZS	1.57	1.74	1.88	1.95	1.99	2.01	2.03	2.03	2.03	

Table 26 - University of Canterbury – Calgary Data Set Concatenated



	512	1,024	2,048	4,096	8,192	16,384	32,768	65,536	131,072	262,144
□XMATCH	2.40	2.52	2.60	2.66	2.69	2.71	2.72	2.72	2.72	2.72
▩SDLT-LZS	2.55	2.55	2.8	2.8	2.8	2.8	2.8	2.8	2.8	

Table 27 - University of Canterbury - Artificial Data Set Concatenated



	512	1024	2048	4096	8192	16384	32768	65536	131072	262144
□XMATCH	1.57	1.73	1.90	2.07	2.21	2.30	2.35	2.37	2.38	2.38
▨SDLT-LZS	1.75	1.91	2.05	2.12	2.15	2.17	2.18	2.19	2.19	

Table 28 - University of Canterbury – "Large" Data Set Concatenated

10. Appendix A – University of Canterbury Data Sets

10.1. Calgary Data Set

File Name	File Description	No. of Files	Size (Kbytes)
Bib	Bibliographic files (refer format)	1	108.65
book1	Hardy: Far from the madding crowd	1	750.75
book2	Witten: Principles of computer speech	1	596.53
Geo	Geophysical data	1	100
News	News batch file	1	368.27
obj1	Compiled code for Vax: compilation of progp	1	21
obj2	Compiled code for Apple Macintosh: Knowledge support system	1	241.02
paper1	Witten, Neal and Cleary: Arithmetic coding for data compression	1	51.91
paper2	Witten: Computer (in)security	1	80.27
paper3 "	Witten: In search of "autonomy	1	45.43
paper4	Cleary: Programming by example revisited	1	12.97
paper5	Cleary: A logical implementation of arithmetic	1	11.67
paper6	Cleary: Compact hash tables using bidirectional linear probing	1	37.21
Pic	Picture number 5 from the CCITT Facsimile test files (text + drawings)	1	501.18
Progc	C source code: compress version 4.0	1	38.68
Progl	Lisp source code: system software	1	69.96
Progp	Pascal source code: prediction by partial matching evaluation program	1	48.22
Trans	Transcript of a session on a terminal	1	91.49
Total		**18**	**3175.21**

Table 29 - Calgary Data Set

10.2. Artificial Data Set

File Name	File Description	No. of Files	Size (Kbytes)
a.txt	The letter 'a'	1	0.001
aaa.txt	The letter 'a', repeated 100,000 times.	1	100
alphabet.txt	Enough repetitions of the alphabet to fill 100,000 characters	1	100
random.txt	100,000 characters, randomly selected from [a-z|A-Z|0-9|!|] (alphabet size 64)	1	100
Total		**4**	**300.001**

Table 30 - Artificial Data Set

10.3. Large Data Set

File Name	File Description	No. of Files	Size (Kbytes)
E.coli	Complete genome of the E. Coli bacterium	1	4639
bible.txt	The King James version of the bible	1	4047
world192.txt	The CIA world fact book	1	2473
Total		**3**	**11159**

Table 31 - Large Data Set