Evaluation of the Term Recognition Task

NTCIR Workshop TMREC Group (Kyo Kageura[†], Masaharu Yoshioka[†], Keita Tsuji[†], Fuyuki Yoshikane[‡], Koichi Takeuchi[†] & Teruo Koyama[†])
[†] Department of Research and Development, NACSIS
{kyo,yoshioka,koichi,koyama}@rd.nacsis.ac.jp
[‡] Graduate School of Education, University of Tokyo
{i34188@m-unix.cc.u-tokyo.ac.jp, fuyuki@p.u-tokyo.ac.jp}

Abstract: In this paper we describe the results of the term recognition task. Result of the comparative analyses of the submitted files vis-a-vis a manually complied list of terms and a list of N-common terms are described. Lastly, the similarity between each file pair is briefly examined.

1 Introduction

In this paper, we describe the results of the term recognition task, which is one of the three subtasks defined by the NTCIR TMREC group. The other two subtasks, i.e. keyword extraction and role analysis, are described separately.

This report is not intended to give an 'evaluation' of the results submitted by participants, though in many places we describe the "performance" of the result files by recall and/or precision. Rather, we aim at clarifying characteristics of the results and methods used in producing the results which can be used to promote discussions related to the task of automatic term recognition. So low scores do not mean inferiority. This is because, unlike IR related tasks, the essential point, i.e. "what is a term?" or "what is terminology?" is a focal point of debate. In addition, the task of term recognition should be evaluated as much by means of internal consistency of the resultant list of terms vis-a-vis the definitions of terms and terminology adopted, as by means of such measures as recall and precision vis-a-vis some pre-compiled lists of candidate terms or possible 'golden standard'. Nevertheless, we analysed the results vis-a-vis the lists of terms prepared by the TMREC group, because for now it is a useful way to clarify the characteristics of result files.

Two lists of terms were prepared by the TMREC group for analysing the results:

a) List of manually extracted term candidates (henceforth Manual-Candidates), and elements listed in the index part of an encyclopaedia in the field of artificial intelligence (Shapiro, 1987) (henceforth Index-Candidates).

b) List of elements constructed from the result files submitted by participants, taking common elements for N files.

We will call the former the 'Candidate evaluation' and latter the 'N-common evaluation'.

In the following, we first give summary descriptions of the submitted files in section 2. In section 3, we describe the result of Candidate evaluation. In section 4, we describe the result of N-common evaluation. In section 5 we observe the similarity of result files.

2 Summary of the Result Files

The number of teams that submitted results in time was 7, with one observer, and the total number of result files was 16, one of which was the observer's. In this paper, we use letters 'a' through 'p' to refer to these 16 result files, to keep them anonymous. Table 1 shows the correspondence between the 8 teams and 16 files. One team, i.e. team 1, submitted a total of 6 files, while the other teams submitted either one or two files.

Table 1 : Correspondence of teams and result files

Team	1	2	3	4	5	6	7	Obs.
Tagged	a–e	g	h	i		l	n	p
Untagged	f			j	k	m	o	

Six teams and the observer used tagged-data. Five files based on tagged-data were submitted by team 1, so in total eleven files were based on tagged-data. Five teams, on the other hand, submitted results based on untagged-data. So the the number of result files based on untagged-data is five.

Among these 16 files, three files, i.e. 'g', 'i', and 'k', were unweighted flat lists of term candidates, while the other 13 files had been given scores indicating the weight of possible 'termhood'.

In addition to the above sixteen files, one file was submitted by a team after the deadline. This file is referred to as 'x'.

2.1 Normalisation of the Result Files

In principle, we did not apply normalisation of the result files. However, after examining the data, we decided to apply the following two normalisations.

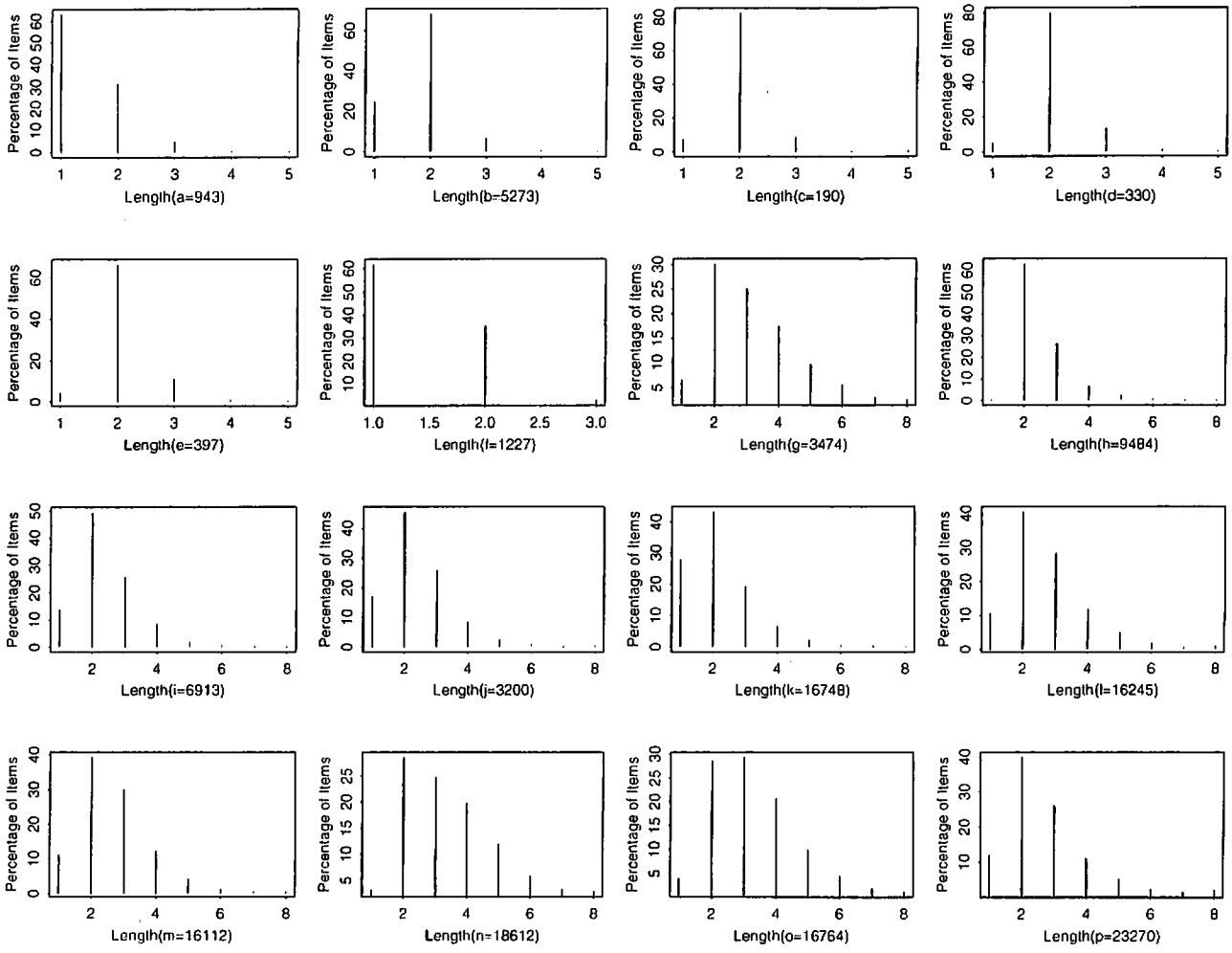


Figure 1 : Distribution of elements by length

- Deletion of one-byte symbols such as {|}, as they are tags in the tagged corpus and have nothing to do with the original language data.
- Deletion of duplicated records. Some seemed to be the result of 'bugs' in the tagged-data, while others seemed to be the result of the methods. We deleted the elements whose weights were lower.

Table 2 shows the basic quantitative data of the result files after these normalisations. All the evaluations and analyses in the following are based on these normalised result files.

Table 2 : Quantities before and after normalisations

File	Before	After	File	Before	After
a	943	943	i	6913	6913
b	5275	5273	j	3200	3200
c	190	190	k	16748	16748
d	330	330	l	16296	16245
e	397	397	m	16112	16112
f	1227	1227	n	18632	18608
g	3474	3474	o	16764	16764
h	9503	9484	p	23308	23270
			x	17865	17757

2.2 Distribution of Candidate Length

In order to clarify the difference between the files, we observed distribution of word length in terms of the number of constituent elements for each file. The calculation was carried out automatically on the basis of the units defined in the tagged-data, so there are some errors originating from the tagged-data. In addition, the elements in the term candidates in the result files based on the untagged-data do not necessarily match the units in the tagged-data. In these cases we made approximations. As a result, for instance, a result file which claims to extract only units with more than two constituent elements includes a few units with only one constituent element. Quantitatively, however, these errors are considered to be negligible.

Table 3 : Summary measures for length (Mo=mode ; Me=median)

	Mean	Mo	Me		Mean	Mo	Me
a	1.44	1	1	i	2.38	2	2
b	1.83	2	2	j	2.37	2	2
c	2.06	2	2	k	2.16	2	2
d	2.13	2	2	l	2.74	2	2
e	2.13	2	2	m	2.70	2	2
f	1.41	1	1	n	3.53	2	3
g	3.34	2	3	o	3.31	3	3
h	2.54	2	2	p	2.82	2	2
				x	2.38	2	2

Table 3 shows the mean (calculated by replacing numbers greater than 8 with 8), mode, and median of the length of candidates in terms of the numbers of constituent elements. Figure 1 shows histograms of the distribution of length in the result files (other than 'x').

In most result files, elements with length 2 are the majority. The files 'b', 'c', 'd', 'e', 'h' are particularly notable with a predominance (more than 60%) of elements with length 2. On the other hand, in the files 'a', 'g', 'n', and 'o', the ratio of elements with length 2 is around 30%. In the other files including 'x', the ratio of length 2 elements is around 40 to 50%.

In the files 'a' and 'f', the ratio of length 1 is the highest. In files 'n' and 'o', the ratio of length 3 is particularly notable; in 'o' the ratio of length 3 is higher than that of length 2. The files 'n', 'o' and to some extent 'g' are notable with respect to the high ratio of longer elements.

As for the median, twelve files, i.e. 'b', 'c', 'd', 'e', 'h', 'i', 'j', 'k', 'l', 'm', 'p' and 'x', take 2. Two files, i.e. 'a' and 'f' take 1, and three files, i.e. 'g', 'n', and 'o', take 3.

Intuitively, we can classify the files into three groups by means of the distribution of the length, i.e. files dominated by elements of length 1 ('a', 'f', ('k')), by elements of length 2 ('b', 'c', 'd', 'e', 'h', 'i', 'j', ('k'), 'l', 'p', 'x'), and by elements of length 3 or more ('g', 'm', 'n', 'o').

If we regard these differences as a result of a deliberate choice of methods, which in turn reflects different viewpoints as to the definition of a 'term', then we can see the range of different conceptualisations of what a 'term' means to the different teams reflected in the files.

3 Candidate-based Evaluation

3.1 Basic Nature of Candidate Terms

As mentioned above, we have prepared two lists of candidate terms, i.e. Manual-Candidates and Index-Candidates. The former constitutes the full superset of the latter. We use 'Candidates' with capital 'C' to refer to these candidate terms prepared by the TMREC group, in order to distinguish them from the term candidates submitted by the participants. This, however, in no way implies that the Candidate lists prepared by the TMREC group have any prescriptive status. There are 8834 Manual-Candidates, and 671 Index-Candidates.

Table 4 : Length of Manual- and Index-Candidates

File	Mean	Mode	Median
Manual-Candidates	2.56	2	2
Index-Candidates	1.76	2	2

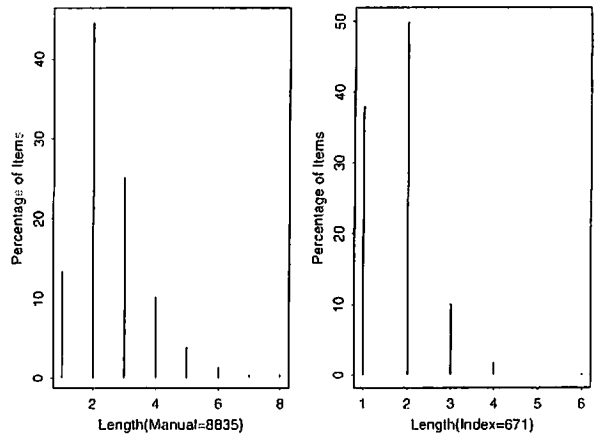


Figure 2 : Distribution of candidates by length

Table 4 shows the distribution of the length of Manual- and Index-Candidates. Figure 2 shows the histogram of the length of the Candidates. In Manual-Candidates, the ratio of length 3 is higher than that of length 1, while in Index-Candidates the ratio of length 1 is higher. However, both seem to be located in the 'center' of the result files shown in Figure 1. As reference lists used to clarify the characteristics of result files, therefore, these Candidates are considered to be useful. Incidentally, Table 13 shows the number of candidates that appear in N files.

3.2 Candidate-based Evaluation (all)

3.2.1 Ignoring Weights

Here we observe the basic nature of result files from the point of view of recall and precision vis-a-vis Candidate lists, without taking into account the weights. Table 5 shows the number of Candidates in each result file. Table 6 lists the precision and recall, which is visually illustrated in Figure 3.

Here let us summarise some notable tendencies. Firstly, we can observe a difference between the results obtained vis-a-vis Manual-Candidates and Index-Candidates. For instance, the files 'a', 'b', 'c', 'd', 'e', and 'f' show high precision with respect to Index-Candidates while they do not show notably high precision with respect to Manual-Candidates.

Table 5: Number of elements matched against Manual- and Index-Candidates

1st char - A : Handmade candidates ; D : Dictionary candidates (AIDI)
2nd char - F : Full match ; A : All inclusive ; I : Result Includes Term Candidates ;
P : Result is a Part of Term Candidates ; B : Result is both I & P

File	Total	AF	AA	AI	AP	AB	DF	DA	DI	DP	DB
a	943	433	810	29	302	46	141	239	51	45	2
b	5275	2351	4237	722	759	405	418	1146	664	54	10
c	190	122	174	6	23	23	28	66	36	0	2
d	330	163	264	28	36	37	34	83	47	0	2
e	397	235	364	16	60	53	49	130	75	2	4
f	1227	566	1032	37	372	57	178	303	72	48	5
g	3474	1252	3170	1778	58	82	155	688	505	22	6
h	9484	3990	7936	3297	150	499	239	2215	1967	4	5
i	6913	3368	5992	2061	410	153	281	1774	1450	39	4
j	3200	1537	2823	764	205	317	162	940	759	16	3
k	16748	6031	13554	4168	2235	1120	548	2998	2378	63	9
l	16245	6377	13710	6000	717	616	359	3306	2924	13	10
m	16112	6536	13177	5464	690	487	356	3372	2996	11	9
n	18608	5554	16673	10484	355	280	449	3041	2526	57	9
o	16764	4013	14596	9580	476	527	357	2358	1941	54	6
p	23270	7944	18711	9432	879	456	640	3934	3215	68	11
x	17757	7330	14155	5100	1263	462	608	3837	3150	68	11

Table 6: Recall and precision based on Manual- and Index-Candidates

1st char - A : Handmade candidates ; D : Dictionary candidates (AIDI)
2nd char - F : Full match ; A : All inclusive ; I : Result Include Term Candidates ;
P : Result is a Part of Term Candidates ; B : Result is both I & P
3rd char - P : Precision ; R : Recall

File	AFP	AAP	AIP	APP	ABP	AFR	DFP	DAP	DIP	DPP	DBP	DFR
a	45.92	85.90	3.08	32.03	4.88	4.90	14.95	25.34	5.41	4.77	0.21	21.01
b	44.59	80.35	13.69	14.39	7.68	26.61	7.93	21.73	12.59	1.02	0.19	62.30
c	64.21	91.58	3.16	12.11	12.11	1.38	14.74	34.74	18.95	0.00	1.05	4.17
d	49.39	80.00	8.48	10.91	11.21	1.84	10.30	25.15	14.24	0.00	0.61	5.07
e	59.19	91.69	4.03	15.11	13.35	2.66	12.34	32.75	18.89	0.50	1.01	7.30
f	46.13	84.11	3.02	30.32	4.65	6.41	14.51	24.69	5.87	3.91	0.41	26.53
g	36.04	91.25	51.18	1.67	2.36	14.17	4.46	19.80	14.54	0.63	0.17	23.10
h	42.07	83.68	34.76	1.58	5.26	45.16	2.52	23.36	20.74	0.04	0.05	35.62
i	48.72	86.68	29.81	5.93	2.21	38.12	4.06	25.66	20.97	0.56	0.06	41.88
j	48.03	88.22	23.88	6.41	9.91	17.40	5.06	29.38	23.72	0.50	0.09	24.14
k	36.01	80.93	24.89	13.34	6.69	68.26	3.27	17.90	14.20	0.38	0.05	81.67
l	39.26	84.40	36.93	4.41	3.79	72.18	2.21	20.35	18.00	0.08	0.06	53.50
m	40.57	81.78	33.91	4.28	3.02	73.98	2.21	20.93	18.59	0.07	0.06	53.06
n	29.85	89.60	56.34	1.91	1.50	62.86	2.41	16.34	13.57	0.31	0.05	66.92
o	23.94	87.07	57.15	2.84	3.14	45.42	2.13	14.07	11.58	0.32	0.04	53.20
p	34.14	80.41	40.53	3.78	1.96	89.92	2.75	16.91	13.82	0.29	0.05	95.38
x	41.28	79.72	28.72	7.11	2.60	82.97	3.42	21.61	17.74	0.38	0.06	90.61

Although we have to take into account the fact that they are small compared to the other result files (except for 'b') and thus have an advantage in obtaining higher precision with respect to a smaller number of Candidates, it can still be argued that, assuming that the Index-Candidates constitute a 'core' set of the terminology, the files 'a'–'f' are inclined to extract the 'core' part of terminology. This tendency can also be observed in file 'j'.

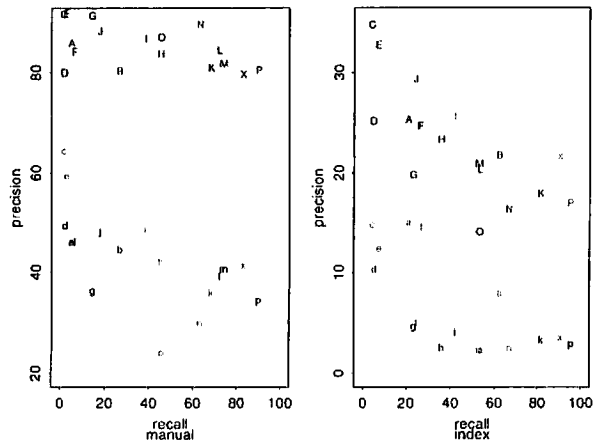
Secondly, the difference between AIP and APP in Table 6 can be noted in the evaluation with respect to Manual-Candidates. AIP refers to precision in which not only a full match, but also the candidates in result files that include Candidates as substrings are counted, while APP refers to precision in which the candidates in result files which are substrings of Candidates are counted. In the files 'a', 'c', 'e', and 'f', the values of APP are higher than those of AIP, while in the files 'g', 'h', 'i', 'j', 'l', 'm', 'n', 'o', 'p', and 'x', the values of AIP are much higher. Especially in 'g', 'h', 'n', and 'o', the difference is notable. In the files 'd' and 'k', the values of AIP and APP are closer.

From the point of view of Manual-Candidates, temporarily assuming they constitute a coherent set of terminology, the methods used to produce 'a', 'c', 'e', and 'f' tend to extract constituent elements of terms, while the methods used to produce the result files 'g', 'h', 'i', 'j', 'l', 'm', 'n', 'o', 'p', and 'x' tend to extract longer units. Among the latter group, 'g', 'n', and 'o', whose full-match precision is low but whose precision AIP is high, seem to be based on different definitions of what 'term' means from the one adopted in Manual-Candidates. The files 'h', 'i', 'j', 'l', and 'm' show the tendency in between these two groups.

Table 7 : Transitions of 11-point precisions

Manual-Candidates														
Recall	a	b	c	d	e	f	h	j	l	m	n	o	p	
0.0	1.0000	0.5104	1.0000	1.0000	0.7361	1.0000	1.0000	1.0000	0.8108	0.8182	1.0000	1.0000	0.9355	
0.1	0.0000	0.4913	0.0000	0.0000	0.0000	0.0000	0.7517	0.5380	0.6852	0.6885	0.3914	0.3272	0.7075	
0.2	0.0000	0.4649	0.0000	0.0000	0.0000	0.0000	0.5877	0.0000	0.6423	0.6453	0.3442	0.2816	0.6212	
0.3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4880	0.0000	0.6098	0.6220	0.3204	0.2661	0.5586	
0.4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4355	0.0000	0.5440	0.5479	0.3085	0.2504	0.5236	
0.5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4458	0.4542	0.3061	0.0000	0.5107	
0.6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4059	0.4148	0.3009	0.0000	0.5034	
0.7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3926	0.4058	0.0000	0.0000	0.4779	
0.8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4507	
0.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
1.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	

Index-Candidates														
Recall	a	b	c	d	e	f	h	j	l	m	n	o	p	
0.0	0.6667	0.1260	0.1667	0.1579	0.2727	1.0000	1.0000	0.1667	0.0641	0.0618	0.3333	0.5000	0.9355	
0.1	0.1909	0.1251	0.0000	0.0000	0.0000	0.1939	0.1949	0.0538	0.0641	0.0618	0.0393	0.0321	0.8590	
0.2	0.1617	0.1251	0.0000	0.0000	0.0000	0.1680	0.0671	0.0538	0.0632	0.0615	0.0336	0.0291	0.8580	
0.3	0.0000	0.1136	0.0000	0.0000	0.0000	0.0000	0.0302	0.0000	0.0601	0.0586	0.0290	0.0262	0.8577	
0.4	0.0000	0.1066	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0521	0.0508	0.0285	0.0247	0.8577	
0.5	0.0000	0.0950	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0226	0.0226	0.0273	0.0219	0.8577	
0.6	0.0000	0.0818	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0252	0.0000	0.8577	
0.7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8504	
0.8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8302	
0.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7665	
1.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	



Lower-case: full-match ;
Upper-case: all including partial match

Figure 3 : Recall and precision (plot of Table 6)

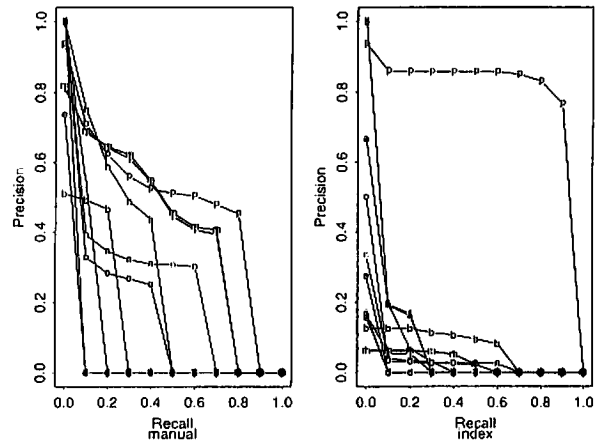


Figure 4 : Transitions of 11-point precisions

With respect to Index-Candidates, characteristics of the relations between the values of DIP and DPP become closer among most files. Other than the files 'a' and 'f', all tend to take longer units with respect to Index-Candidates.

3.2.2 Considering Weights

Let us here take into consideration the weights given to term candidates. We only observe tendencies on the basis of a full match. Figure 4 shows simple 11-point scores applied to the result files with weights, i.e. 13 files other than 'g', 'i', 'k' and 'x'[1], matched against Manual- and Index-Candidates. Table 7 shows the same information by scores.

The horizontal axis in Figure 4 is recall, so it totally depends on the recall (AFR and DFR) in Table 6, because 11-point scores automatically become 0 when the horizontal axis goes beyond the original recall of the result files. For instance, for file 'l', whose recall with respect to the Manual-Candidates is over 70%, 11-point values up to that interval take positive values, then become 0 beyond that interval. The file 'a', on the other hand, whose overall recall is less than 5%, takes 0 for 10 out of 11-points. So, we only show the transition patterns but not the 11-point average, because the average scores do not mean anything. Transition patterns are considered to be useful to clarify the characteristics of result files.

For instance, it might be useful to examine the different results obtained by using the same methods with a different corpus, i.e. a tagged and an untagged corpus (see for instance 'l' and 'm' in contrast with 'n' and 'o').

The file 'p' shows the highest transition pattern with respect to Index-Candidates. This is partly because the number of candidate terms in 'p' is very large, and also because the ATR method used in producing 'p' utilises the Index-Candidates as a part of the information sources (though not in simple match-

[1] term candidates in the file 'x' are with scores, but they correspond to the specificity of concepts and are thus excluded from here.

Table 8 : Evaluation by intervals (up to 3000 in intervals of 1000)

Manual-Candidates										
Interval	a	b	f	h	j	l	m	n	o	p
~1000	0.4330	0.4930	0.4890	0.7650	0.5670	0.7010	0.7050	0.4490	0.3890	0.7440
~2000	0	0.4830	0	0.6040	0.4430	0.6030	0.6070	0.3520	0.2950	0.5840
~3000	0	0.4280	0	0.3940	0.4420	0.6180	0.6180	0.3270	0.2740	0.5180
Index-Candidates										
Interval	a	b	f	h	j	l	m	n	o	p
~1000	0.1410	0.1210	0.1610	0.1040	0.0430	0.0550	0.0530	0.0480	0.0440	0.6110
~2000	0	0.1030	0	0.0290	0.0480	0.0490	0.0450	0.0280	0.0220	0.0030
~3000	0	0.0750	0	0.0160	0.0630	0.0790	0.0830	0.0320	0.0250	0.0040

Table 9 : Evaluation by intervals (up to 15000 in intervals of 3000)

Manual-Candidates								
Interval	b	h	j	l	m	n	o	p
~3000	0.4680	0.5877	0.4840	0.6407	0.6433	0.3760	0.3193	0.6153
~6000	0.3157	0.3670	0	0.4940	0.4993	0.2977	0.2500	0.4490
~9000	0	0.3217	0	0.2623	0.2547	0.2760	0.2323	0.4603
~12000	0	0	0	0.2587	0.2843	0.2787	0.2357	0.4387
~15000	0	0	0	0.3023	0.3337	0.2917	0.1987	0.3113
Index-Candidates								
Interval	b	h	j	l	m	n	o	p
~3000	0.0997	0.0497	0.0513	0.0610	0.0603	0.0360	0.0303	0.2060
~6000	0.0397	0.0147	0	0.0363	0.0353	0.0217	0.0230	0.0030
~9000	0	0.0117	0	0.0067	0.0067	0.0280	0.0240	0.0010
~12000	0	0	0	0.0020	0.0027	0.0233	0.0180	0.0020
~15000	0	0	0	0.0037	0.0037	0.0193	0.0140	0

ing).

We also observed the precision of some fixed intervals, using 1000 and 3000 as the interval sizes. By using 1000 as an interval, we could take at least one interval for 'a' and 'f' (but we excluded 'c', 'd', 'e', submitted by the same team). By using 3000, we could trace the behaviour of the bigger result files while still taking two intervals for middle sized 'b' and 'i'.

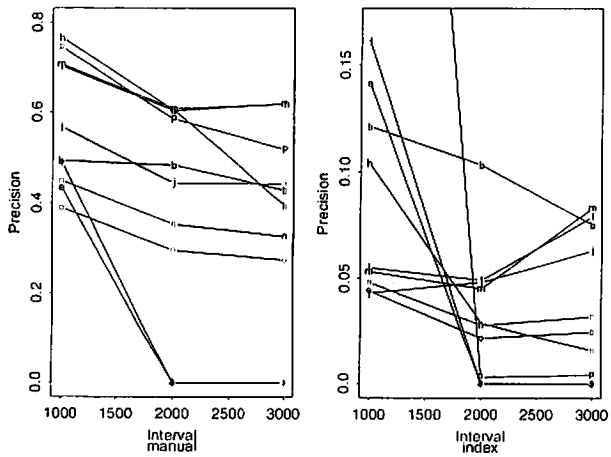


Figure 5 : Evaluation by intervals (up to 3000 in intervals of 1000)

Here we observed the precisions of intervals 1–1000, 1001–2000, etc. instead of observing the precisions of cumulative intervals 1–1000, 1–2000, etc. In the evaluation of IR, the latter method is usually adopted. This seems to be because in IR it is assumed that the main purpose is to evaluate the results and also that users are assumed to examine the results from the top. However, for diagnostic purposes, it is better to observe the intervals 1–1000, 1001–2000, etc. This is why we adopted this method.

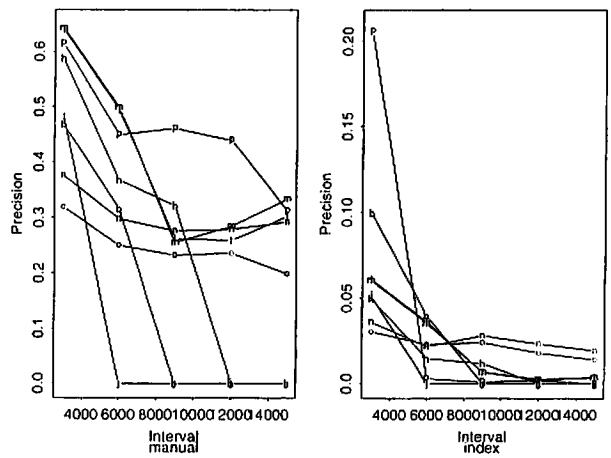


Figure 6 : Evaluation by intervals (up to 15000 in intervals of 3000)

Table 8 and Figure 5 show the results of precisions up to 3000 with intervals of 1000 (3 intervals). Table 9 and Figure 5 shows the results up to 15000 with

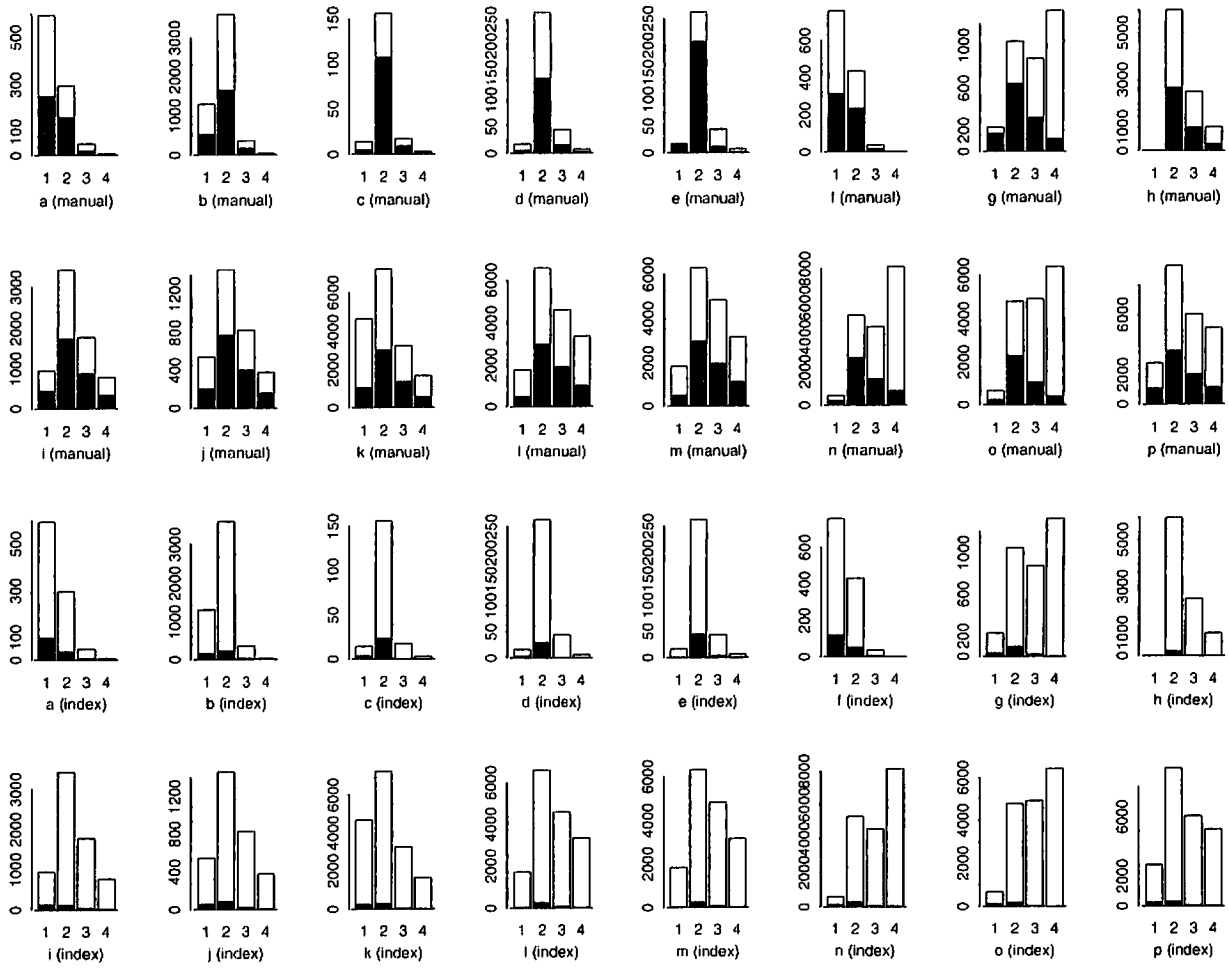


Figure 7 : Distribution of matched elements by length

intervals of 3000 (5 intervals).

A successful method or system which adopts the same concepts of 'terms' and 'terminology' as Manual- or Index-Candidates must show a downward curve from left to right. From this, we can intuitively argue that result files which show a downward curve from left to right have adopted the same type of conceptualisation of 'terms' and 'terminology'. On the other hand, if result files show an upward curve from left to right, this either means the system performance is not good or the system is based on different conceptualisations of 'terms' and 'terminology' from those adopted in Manual- or Index-Candidates.

From this point of view, it is interesting to observe (excluding 'p', 'a' and 'f') that the files 'b' and 'h' show clear downward curves from left to right against Index-Candidates, while other files show irregular transitions. The definition of 'terms' in 'b' and 'h' may be similar to that in Index-Candidates.

From Figure 6, with respect to Manual-Candidates (which ignores very small files, but the very fact that the result files are small implies that these small files adopt different definitions of 'terms' and 'terminology' from that of Manual-Candidates), it is notable that the files 'n' and 'o' show stable values from the first interval to the last interval. This may well show that these result files are based on different conceptualisations of 'terms' and 'terminology' from those adopted by Manual-Candidates. 'l' and 'm' to some extent show similar tendencies, from the mid-interval. These files have a different distribution of the length of candidates, which corresponds to the tendencies observed here.

3.3 Candidate-based Evaluation (by length)

Figure 7 shows the number of candidates in each file (except 'x') which match the Manual- and Index-Candidates by length (1, 2, 3, and 4-plus).

Tables 10 and 11 show the number of candidates matched against Manual- and Index-Candidates and the precision and recall respectively, by length. Figures 8 and 9 show the transitions of recall (left panel) and precision (right panel) by length, against Manual- and Index-Candidates respectively (here 'x' is excluded). Figures 10 and 11 show the relations of

Table 11 : Recall and precision by length
(2nd char: P — precision, R — recall ; 3rd char: length)

File	AP1	AR1	AP2	AR2	AP3	AR3	AP4	AR4	DP1	DR1	DP2	DR2	DP3	DR3	DP4	DR4
a	0.42	0.21	0.55	0.04	0.42	0.01	0.29	0.00	0.16	0.37	0.13	0.11	0.17	0.12	0.00	0.00
b	0.40	0.43	0.46	0.42	0.48	0.08	0.50	0.01	0.12	0.60	0.07	0.70	0.08	0.40	0.09	0.23
c	0.36	0.00	0.69	0.03	0.53	0.00	0.33	0.00	0.00	0.00	0.15	0.07	0.24	0.06	0.00	0.00
d	0.29	0.00	0.53	0.04	0.36	0.01	0.43	0.00	0.00	0.00	0.11	0.09	0.11	0.07	0.00	0.00
e	0.43	0.01	0.62	0.05	0.41	0.00	1.00	0.00	0.03	0.00	0.13	0.13	0.11	0.04	0.00	0.00
f	0.42	0.26	0.54	0.06	0.43	0.01	0.00	0.00	0.16	0.47	0.12	0.15	0.19	0.10	0.00	0.00
g	0.75	0.14	0.61	0.16	0.36	0.14	0.09	0.08	0.15	0.13	0.09	0.28	0.03	0.35	0.00	0.15
h	0.50	0.01	0.45	0.68	0.39	0.44	0.30	0.20	0.08	0.01	0.04	0.62	0.01	0.37	0.00	0.23
i	0.46	0.36	0.50	0.44	0.49	0.39	0.44	0.23	0.13	0.49	0.04	0.36	0.02	0.46	0.01	0.38
j	0.37	0.17	0.53	0.20	0.49	0.18	0.42	0.11	0.10	0.21	0.06	0.25	0.02	0.28	0.01	0.38
k	0.23	0.00	0.42	0.77	0.42	0.61	0.35	0.39	0.05	0.91	0.04	0.78	0.02	0.76	0.00	0.31
l	0.28	0.40	0.46	0.76	0.41	0.86	0.30	0.68	0.03	0.18	0.04	0.74	0.01	0.87	0.00	0.54
m	0.27	0.40	0.47	0.76	0.41	0.88	0.35	0.76	0.02	0.18	0.04	0.73	0.01	0.85	0.00	0.54
n	0.51	0.25	0.53	0.71	0.35	0.71	0.11	0.58	0.21	0.49	0.05	0.79	0.01	0.81	0.00	0.54
o	0.39	0.22	0.48	0.58	0.22	0.48	0.06	0.28	0.17	0.45	0.04	0.60	0.01	0.57	0.00	0.31
p	0.39	0.91	0.39	0.93	0.34	0.92	0.23	0.79	0.09	0.96	0.03	0.95	0.01	0.97	0.00	0.92
0.27	0.76	0.47	0.90	0.46	0.85	0.37	0.66	0.07	0.95	0.04	0.89	0.01	0.90	0.00	0.54	

Table 10 : Number of matched elements by length
A1 : Against Manual-Candidates, length 1,
D1 : Against Index-Candidates, length 1, etc

File	A1	A2	A3	A4	D1	D2	D3	D4
a	250	161	20	2	95	38	8	0
b	516	1647	171	17	153	235	27	3
c	5	107	9	1	24	4	0	0
d	5	139	16	3	29	5	0	0
e	15	208	11	1	1	45	3	0
f	315	235	16	0	121	50	7	0
g	171	640	319	122	34	95	24	2
h	13	2691	983	302	2	209	25	3
i	435	1726	870	337	124	121	31	5
j	203	773	401	160	53	85	19	5
k	1071	3037	1348	574	231	261	52	4
l	476	3000	1900	1001	45	248	59	7
m	483	2981	1958	1114	45	246	58	7
n	295	2813	1587	859	124	263	55	7
o	260	2287	1058	408	114	200	39	4
p	1085	3652	2038	1169	245	317	66	12
x	911	3558	1885	976	242	298	61	7

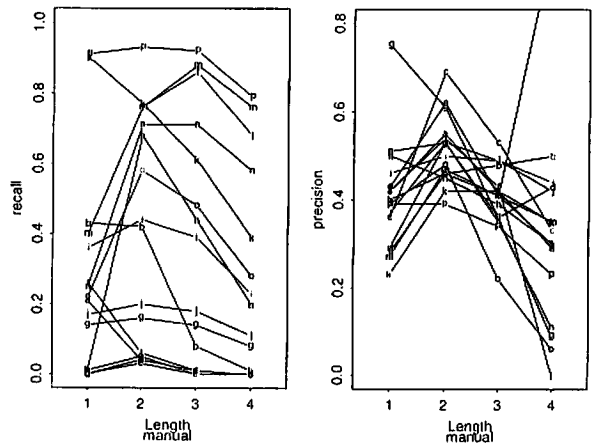


Figure 8 : Recall and precision by length for Manual-Candidates

precision and recall for each length 1, 2, 3, 4-plus, against Manual- and Index-Candidates respectively.

From Figures 8 and 9, we can observe a few characteristic result files. With respect to recall, we can observe that 'a', 'b', 'f', and 'k' perform relatively well in shorter elements compared to longer elements. On the other hand, 'l', 'm' and 'n' show relatively good performance for longer elements. Also, 'h' and 'o' are notable in the gap between performances with length 2 and with other lengths. With respect to precision, 'g' is notable in that it shows a steep downward curve from left to right.

With respect to Manual-Candidates, most result files take maximum values in length 2, but with respect to Index-Candidates the pattern is slightly different. For instance, the precisions of 'n' and 'o' for length 1 are notably high (though the recall is low), while 'a', 'f', and 'c' take high values for length 3.

Similar tendencies can be observed from Figures 10 and 11. 'j' and 'p' show stability with respect to recall, while 'i' and 'b' show stability with respect to precision.

4 N-Common Evaluation

The N-common evaluation is based on the idea that elements that appear in more result files are more likely to be proper terms. Here we first briefly summarise the nature of N-common data, and then observe the characteristics of result files with respect to N-common data.

4.1 N-Common Data

We produced three types of N-common data. The first, henceforth 'Ca', was produced by 13 files (excluding 'c', 'd', 'e' and 'x'). The second, 'C1' was produced by 8 files based on tagged-data (excluding 'c', 'd', 'e' and 'x'), and the third and last, 'C2' was produced by 5 files based on untagged-data. The files 'c', 'd', and 'e' were excluded because they were submitted by the same team as 'a', 'b' and 'f'. The file 'x' was not included because N-common data was made before the submission of 'x'. We used these three N-common files in our evaluation. It would be expected that the teams that submitted two files (from tagged- and untagged-data) would obtain relatively better score if 'Ca' was used. However, we used 'Ca' anyway as the effects can be cancelled out by proper observation of 'C1' and 'C2'. Once again we want to emphasise that our main purpose is to clarify the characteristics of result files, and not to evaluate them

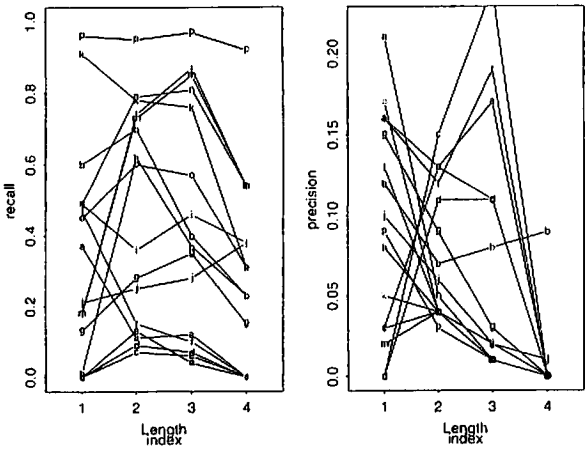


Figure 9 : Recall and precision by length for Index-Candidates

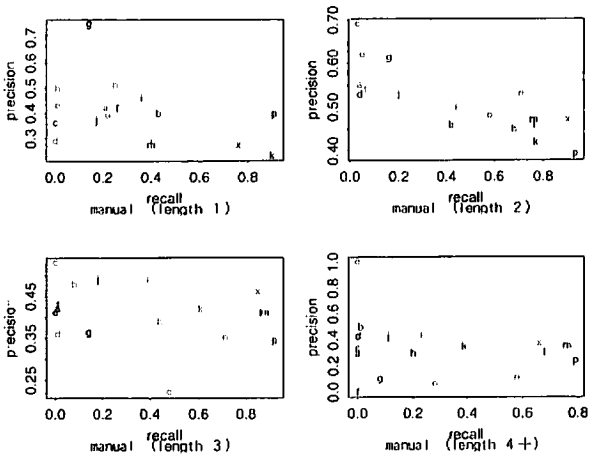


Figure 10 : Recall and precision by length for Manual-Candidates

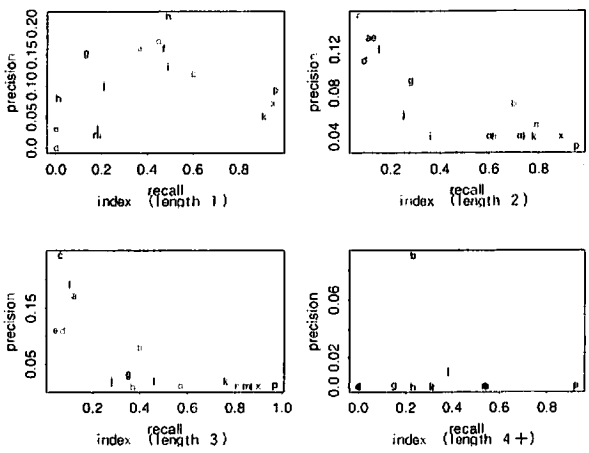


Figure 11 : Recall and precision by length for Index-Candidates

or determine which result files are better than others. Table 12 shows the number of N-commons for each N.

Table 12 : Frequency by N in N-common data

Ca		C1		C2	
n	Freq	n	Freq	n	Freq
1	20631	1	19201	1	21042
2	7997	2	4940	2	7254
3	3712	3	4639	3	4672
4	2952	4	4708	4	1030
5	2919	5	3106	5	73
6	3030	6	933	-	-
7	2576	7	156	-	-
8	2024	8	20	-	-
9	885	-	-	-	-
10	226	-	-	-	-
11	77	-	-	-	-
12	42	-	-	-	-
13	9	-	-	-	-
Total	47080	Total	37703	Total	34071

Table 13 : Manual- and Index-Candidates and N-common data

	Manual		Index	
n	Freq	Recall	Freq	Recall
0	313	-	8	-
1	329	0.02	13	0.00
2	417	0.05	39	0.00
3	558	0.15	45	0.01
4	1049	0.36	69	0.02
5	1184	0.41	68	0.02
6	1432	0.47	76	0.03
7	1408	0.55	104	0.04
8	1232	0.61	105	0.05
9	599	0.68	78	0.09
10	196	0.87	30	0.13
11	69	0.90	19	0.25
12	39	0.93	15	0.36
13	9	1.00	2	0.22

Table 13 shows the number and ratio of N in 'Ca' in Manual- and Index-Candidates, for information. Note however that the analyses here are independent of the analyses in the previous section carried out on the basis of Manual- and Index-Candidates.

We used N-common data in two ways. Firstly, we observed the ratio of N as a whole. Secondly, we observed recall and precision assuming that N-common data with N larger than a certain value are term Candidates. We refer to these as 'Cx-N-Candidates', e.g. Ca-8-Candidates, C1-4-Candidates, etc. For instance, Ca-8-Candidates means that candidates with 8- or more commons in Ca are considered to be the Candidates of terms.

4.2 Ratio of N

Table 14 shows the ratio of N for each N. Figures 12 and 13 show the number of N for larger N (Ca

		Ca		C1		C2			Ca		C1		C2			
file	N	num	ratio	num	ratio	num	ratio	file	N	num	ratio	num	ratio	num	ratio
a	1	11	1.17	56	5.94	137	14.53	b	1	615	11.66	888	16.84	1399	26.53
a	2	38	4.03	140	14.85	264	28.00	b	2	446	8.46	629	11.93	1373	26.04
a	3	81	8.59	146	15.48	316	33.51	b	3	449	8.52	891	16.90	1109	21.03
a	4	75	7.95	252	26.72	140	14.85	b	4	582	11.04	1011	19.17	333	6.32
a	5	103	10.92	162	17.18	51	5.41	b	5	549	10.41	1052	19.95	70	1.33
a	6	156	16.54	93	9.86	0	0.00	b	6	635	12.04	627	11.89	0	0.00
a	7	151	16.01	74	7.85	0	0.00	b	7	641	12.16	155	2.94	0	0.00
a	8	118	12.51	20	2.12	0	0.00	b	8	628	11.91	20	0.38	0	0.00
a	9	73	7.74	0	0.00	0	0.00	b	9	431	8.17	0	0.00	0	0.00
a	10	45	4.77	0	0.00	0	0.00	b	10	170	3.22	0	0.00	0	0.00
a	11	46	4.88	0	0.00	0	0.00	b	11	76	1.44	0	0.00	0	0.00
a	12	37	3.92	0	0.00	0	0.00	b	12	42	0.80	0	0.00	0	0.00
a	13	9	0.95	0	0.00	0	0.00	b	13	9	0.17	0	0.00	0	0.00
c	1	4	2.11	8	4.21	32	16.84	d	1	7	2.12	18	5.45	77	23.33
c	2	8	4.21	27	14.21	33	17.37	d	2	15	4.55	59	17.88	65	19.70
c	3	19	10.00	6	3.16	30	15.79	d	3	41	12.42	21	6.36	58	17.58
c	4	8	4.21	26	13.68	48	25.26	d	4	25	7.58	57	17.27	70	21.21
c	5	6	3.16	28	14.74	36	18.95	d	5	19	5.76	51	15.45	42	12.73
c	6	14	7.37	32	16.84	0	0.00	d	6	34	10.30	48	14.55	0	0.00
c	7	18	9.47	51	26.84	0	0.00	d	7	28	8.48	60	18.18	0	0.00
c	8	17	8.95	12	6.32	0	0.00	d	8	34	10.30	15	4.55	0	0.00
c	9	17	8.95	0	0.00	0	0.00	d	9	25	7.58	0	0.00	0	0.00
c	10	17	8.95	0	0.00	0	0.00	d	10	27	8.18	0	0.00	0	0.00
c	11	31	16.32	0	0.00	0	0.00	d	11	38	11.52	0	0.00	0	0.00
c	12	23	12.11	0	0.00	0	0.00	d	12	27	8.18	0	0.00	0	0.00
c	13	8	4.21	0	0.00	0	0.00	d	13	9	2.73	0	0.00	0	0.00
e	1	28	7.05	44	11.08	81	20.40	f	1	38	3.10	125	10.19	173	14.10
e	2	27	6.80	33	8.31	74	18.64	f	2	84	6.85	174	14.18	411	33.50
e	3	29	7.30	41	10.33	70	17.63	f	3	105	8.56	192	15.65	399	32.52
e	4	20	5.04	55	13.85	66	16.62	f	4	129	10.51	267	21.76	171	13.94
e	5	27	6.80	59	14.86	33	8.31	f	5	140	11.41	181	14.75	73	5.95
e	6	32	8.06	62	15.62	0	0.00	f	6	171	13.94	115	9.37	0	0.00
e	7	42	10.58	50	12.59	0	0.00	f	7	175	14.26	76	6.19	0	0.00
e	8	34	8.56	9	2.27	0	0.00	f	8	137	11.17	19	1.55	0	0.00
e	9	34	8.56	0	0.00	0	0.00	f	9	89	7.25	0	0.00	0	0.00
e	10	29	7.30	0	0.00	0	0.00	f	10	57	4.65	0	0.00	0	0.00
e	11	31	7.81	0	0.00	0	0.00	f	11	51	4.16	0	0.00	0	0.00
e	12	20	5.04	0	0.00	0	0.00	f	12	42	3.42	0	0.00	0	0.00
e	13	7	1.76	0	0.00	0	0.00	f	13	9	0.73	0	0.00	0	0.00
g	1	191	5.50	400	11.77	1180	33.97	h	1	1023	10.79	1404	14.80	1665	17.56
g	2	554	15.95	1030	29.65	592	17.04	h	2	468	4.93	569	6.00	2845	30.00
g	3	688	19.80	175	5.04	775	22.31	h	3	478	5.04	1197	12.62	2829	29.83
g	4	125	3.60	387	11.14	303	8.72	h	4	539	5.68	2720	28.68	706	7.44
g	5	175	5.04	735	21.16	41	1.18	h	5	1061	11.19	2580	27.20	50	0.53
g	6	242	6.97	583	16.78	0	0.00	h	6	1477	15.57	851	8.97	0	0.00
g	7	354	10.19	135	3.89	0	0.00	h	7	1685	17.77	143	1.51	0	0.00
g	8	457	13.15	20	0.58	0	0.00	h	8	1664	17.55	20	0.21	0	0.00
g	9	429	12.35	0	0.00	0	0.00	h	9	780	8.22	0	0.00	0	0.00
g	10	162	4.66	0	0.00	0	0.00	h	10	199	2.10	0	0.00	0	0.00
g	11	52	1.50	0	0.00	0	0.00	h	11	66	0.70	0	0.00	0	0.00
g	12	36	1.04	0	0.00	0	0.00	h	12	35	0.37	0	0.00	0	0.00
g	13	9	0.26	0	0.00	0	0.00	h	13	0	0.00	0	0.00	0	0.00
i	1	8	0.12	11	0.16	1525	22.06	j	1	562	17.56	292	9.12	706	22.06
i	2	69	1.00	392	5.67	2571	37.19	j	2	192	6.00	294	9.19	564	17.62
i	3	311	4.50	1226	17.73	2107	30.48	j	3	166	5.19	447	13.97	901	28.16
i	4	656	9.49	2212	32.00	479	6.93	j	4	192	6.00	622	19.44	956	29.88
i	5	1012	14.64	2188	31.65	38	0.55	j	5	231	7.22	564	17.62	73	2.28
i	6	1423	20.58	739	10.69	0	0.00	j	6	304	9.50	236	7.38	0	0.00
i	7	1222	17.68	125	1.81	0	0.00	j	7	409	12.78	78	2.44	0	0.00
i	8	1320	19.09	20	0.29	0	0.00	j	8	457	14.28	12	0.38	0	0.00
i	9	647	9.36	0	0.00	0	0.00	j	9	420	13.12	0	0.00	0	0.00
i	10	155	2.24	0	0.00	0	0.00	j	10	166	5.19	0	0.00	0	0.00
i	11	53	0.77	0	0.00	0	0.00	j	11	58	1.81	0	0.00	0	0.00
i	12	28	0.41	0	0.00	0	0.00	j	12	34	1.06	0	0.00	0	0.00
i	13	9	0.13	0	0.00	0	0.00	j	13	9	0.28	0	0.00	0	0.00
k	1	2986	17.83	1935	11.55	5367	32.05	l	1	1715	10.56	2964	18.25	3418	21.04
k	2	1231	7.35	1875	11.20	5810	34.69	l	2	916	5.64	2202	13.55	5435	33.46
k	3	1516	9.05	2850	17.02	4476	26.73	l	3	1585	9.76	3421	21.06	4183	25.75
k	4	1545	9.22	3487	20.82	1022	6.10	l	4	1979	12.18	3890	23.95	959	5.90
k	5	1908	11.39	2462	14.70	73	0.44	l	5	2284	14.06	2755	16.96	73	0.45
k	6	2381	14.22	761	4.54	0	0.00	l	6	2475	15.24	845	5.20	0	0.00
k	7	2147	12.82	134	0.80	0	0.00	l	7	2243	13.81	148	0.91	0	0.00
k	8	1850	11.05	18	0.11	0	0.00	l	8	1875	11.54	20	0.12	0	0.00
k	9	841	5.02	0	0.00	0	0.00	l	9	835	5.14	0	0.00	0	0.00
k	10	222	1.33	0	0.00	0	0.00	l	10	212	1.31	0	0.00	0	0.00
k	11	72	0.43	0	0.00	0	0.00	l	11	75	0.46	0	0.00	0	0.00
k	12	40	0.24	0	0.00	0	0.00	l	12	42	0.26	0	0.00	0	0.00
k	13	9	0.05	0	0.00	0	0.00	l	13	9	0.06	0	0.00	0	0.00
m	1	1722	10.69	1469	9.12	5345	33.17	n	1	2704	14.53	6056	37.38	6952	37.36
m	2	797	4.95	1914	11.88	5469	33.94	n	2	4677	25.13	1613	8.67	3280	17.63
m	3	1560	9.68	3394	21.07	4256	26.42	n	3	1162	6.24	2320	12.47	3976	21.37
m	4	1967	12.21	3844	23.86	969	6.01	n	4	967	5.20	3661	19.67	953	5.12
m	5	2258	14.01	2732	16.96	73	0.45	n	5	1466	7.88	2955	15.88	70	0.38
m	6	2498	15.50	844	5.24	0	0.00	n	6	2139	11.50	927	4.98	0	0.00
m	7	2260	14.03	146	0.91	0	0.00	n	7	2303	12.38	156	0.84	0	0.00
m	8	1880	11.67	20	0.12	0	0.00	n	8	1969	10.58	20	0.11	0	0.00
m	9	834	5.18	0	0.00	0	0.00	n	9	870	4.68	0	0.00	0	0.00
m	10	211	1.31	0	0.00	0	0.00	n	10	223	1.20	0	0.00	0	0.00
m	11	74	0.46	0	0.00	0	0.00	n	11	77	0.41	0	0.00	0	0.00
m	12	42	0.26	0	0.00	0	0.00	n	12	42	0.23	0	0.00	0	0.00
m	13	9	0.06	0	0.00	0	0.00	n	13	9	0.05	0	0.00	0	0.00
o	1	3791	22.61	4977	29.69	9451	56.38	p	1	5265	22.63	6513	27.99	5449	23.42
o	2	4848	28.92	1200	7.16	2254	13.45	p	2	1674	7.19	3305	14.20	5965	25.63
o	3	1074	6.41	1207	7.20	3984	23.77	p	3	1961	8.43	4541	19.51	4550	19.55
o	4	375	2.24	2254	13.45	1002	5.98	p	4	2677	11.50	4699	20.19	1021	4.39
o	5	558	3.33	2255	13.45	73	0.44	p	5	2850	12.25	3103	13.33	73	0.31
o	6	1265	7.55	761	4.54	0	0.00	p	6	3014	12.95	933	4.01	0	0.00
o	7	1876	11.19	132	0.79	0	0.00	p	7	2566	11.03	156	0.67	0	0.00
o	8	1813	10.81	19	0.11	0	0.00	p	8	2024	8.70	20	0.09	0	0.00
o	9	831	4.96	0	0.00	0	0.00	p	9	885	3.80	0	0.00	0	0.00
o	10	212	1.26	0	0.00	0	0.00	p	10	226	0.97	0	0.00	0	0.00
o	11	70	0.42	0	0.00	0	0.00	p	11	77	0.33	0	0.00	0	0.00
o	12	42	0.25	0	0.00	0	0.00	p	12	42	0.18	0	0.00	0	0.00
o	13	9	0.05	0	0.00	0	0.00	p	13	9	0.04	0	0.00	0	0.00

Table 14 : Ratio of N-common

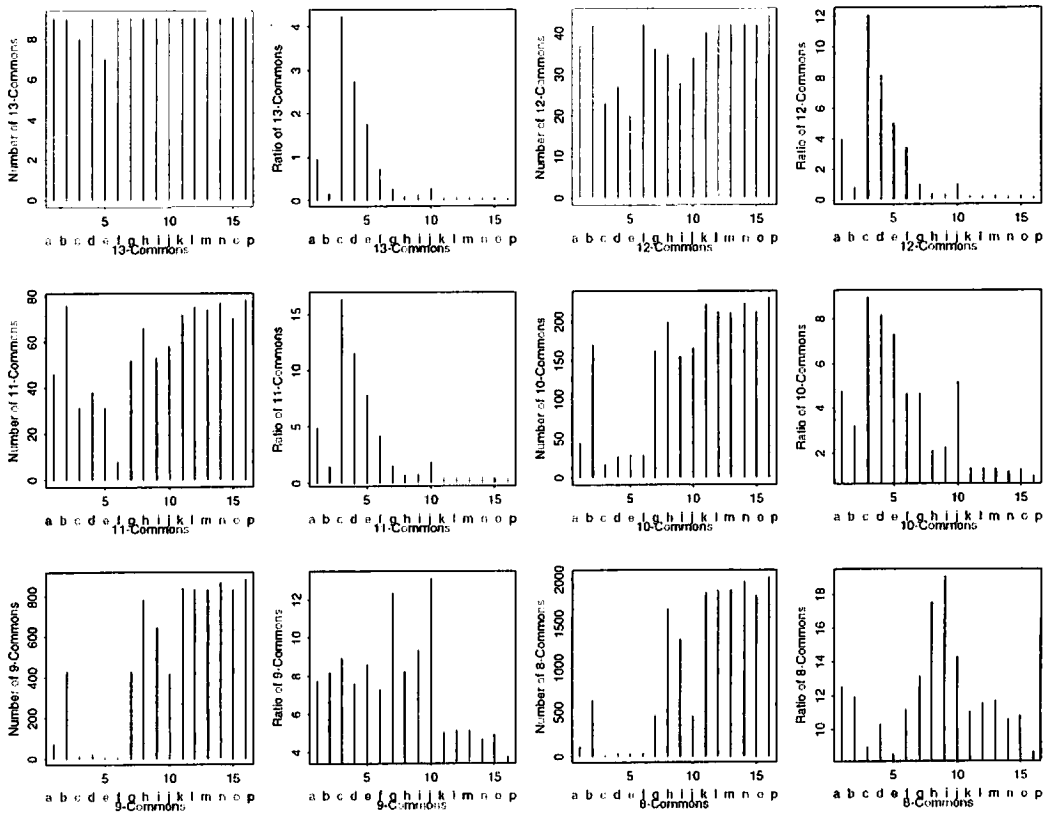


Figure 12 : Number of match against Ca-N-common

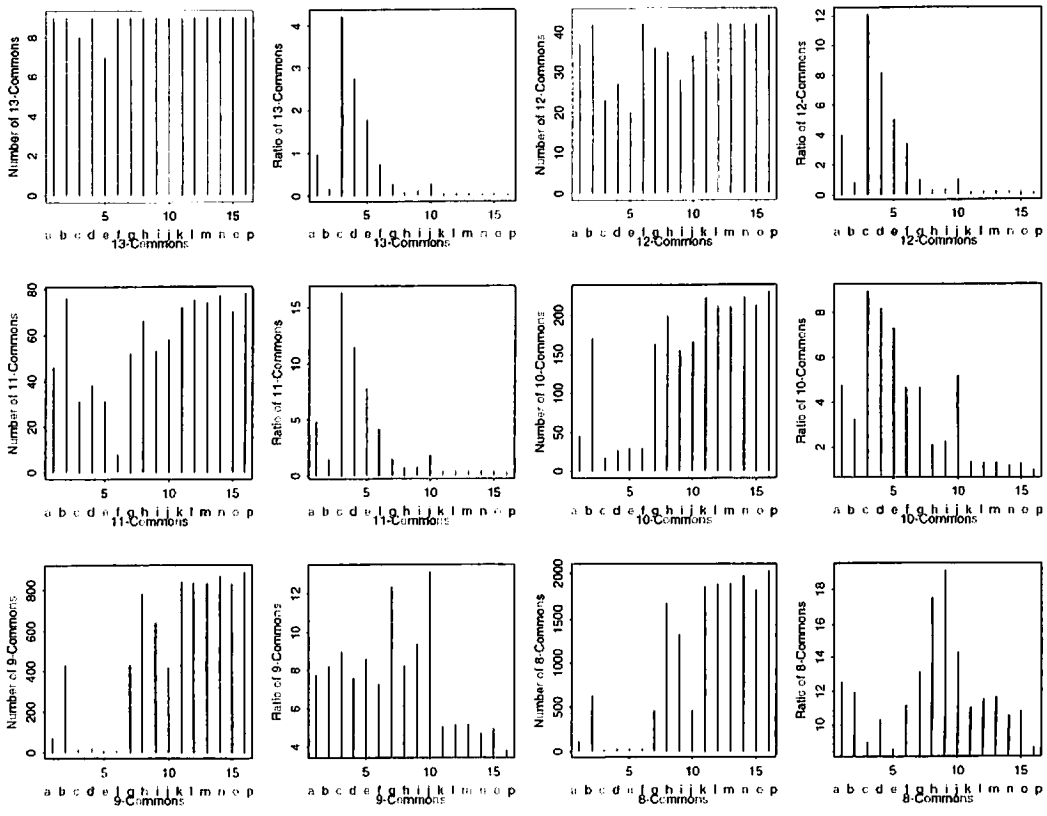


Figure 13 : Number of match against C1-(top 2 rows) and C2-(bottom row) N-common

Table 15 : Frequency of matched elements against N-common data

file	Ca 10+	Ca 9+	Ca 8+	Ca 7+	Ca 6+	Ca 5+	C1 7+	C1 6+	C1 5+	C1 4+	C2 4+	C2 3+
a	137	210	328	479	635	738	94	187	349	601	191	507
b	297	728	1356	1997	2632	3181	175	802	1854	2865	403	1512
c	79	96	113	131	145	151	63	95	123	149	84	114
d	101	126	160	188	222	241	75	123	174	231	112	170
e	87	121	155	197	220	256	59	121	180	235	99	169
f	159	248	385	560	731	871	95	210	391	658	244	643
g	259	688	1145	1499	1741	1916	155	738	1473	1860	344	1119
h	309	1089	2753	4438	5915	6976	163	1014	3594	6314	756	3585
i	245	892	2212	3434	4857	5869	145	884	3072	5284	517	2624
j	267	687	1144	1553	1857	2088	90	326	890	1512	1029	1930
k	343	1184	3034	5181	7562	9470	152	913	3375	6862	1095	5571
l	338	1173	3048	5291	7766	10050	168	1013	3768	7658	1032	5215
m	336	1170	3050	5310	7808	10066	166	1010	3742	7586	1042	5298
n	351	1221	3190	5493	7632	9098	176	1103	4058	7719	1023	4999
o	333	1164	2977	4853	6118	6676	151	912	3167	5421	1075	5059
p	354	1239	3263	5829	8843	11693	176	1109	4212	8011	1094	5644
x	354	1239	3257	5814	8767	11401	176	1109	4185	8738	1094	5610

Table 16 : Recall and precision against N-common data

file		Ca 10+	Ca 9+	Ca 8+	Ca 7+	Ca 6+	Ca 5+	C1 7+	C1 6+	C1 5+	C1 4+	C2 4+	C2 3+
a	pre	14.53	22.27	34.78	50.80	67.34	78.26	9.97	19.83	37.01	63.73	20.25	53.76
	rec	38.70	16.95	10.05	8.20	7.16	6.26	53.41	16.86	8.28	6.74	17.32	8.78
b	pre	5.63	13.81	25.72	37.87	49.91	60.33	3.32	15.21	35.16	54.33	7.64	28.67
	rec	83.90	58.76	41.56	34.20	29.68	26.99	99.43	72.32	43.99	32.11	36.54	26.18
c	pre	41.58	50.53	59.47	68.95	76.32	79.47	33.16	50.00	64.74	78.42	44.21	60.00
	rec	22.32	7.75	3.46	2.24	1.63	1.28	35.80	8.57	2.92	1.67	7.62	1.97
d	pre	30.61	38.18	48.48	56.97	67.27	73.03	22.73	37.27	52.73	70.00	33.94	51.52
	rec	28.53	10.17	4.90	3.22	2.50	2.04	42.61	11.09	4.13	2.59	10.15	2.94
e	pre	21.91	30.48	39.04	49.62	57.68	64.48	14.86	30.48	45.34	59.19	24.94	42.57
	rec	24.58	9.77	4.75	3.37	2.58	2.17	33.52	10.91	4.27	2.63	8.98	2.93
f	pre	12.96	20.21	31.38	45.64	59.58	70.99	7.74	17.11	31.87	53.63	19.89	52.40
	rec	44.92	20.02	11.80	9.59	8.24	7.39	53.98	18.04	9.28	7.37	22.12	11.13
g	pre	7.46	19.80	32.96	43.15	50.12	55.15	4.46	21.24	42.40	53.54	9.90	32.21
	rec	73.16	55.53	35.09	25.67	19.63	16.25	88.07	66.55	34.95	20.85	31.19	19.38
h	pre	3.26	11.48	29.03	46.79	62.37	73.56	1.72	10.69	37.90	66.58	7.97	37.80
	rec	87.29	87.89	84.37	76.01	66.69	59.18	92.61	91.43	85.27	70.76	68.54	62.08
i	pre	3.54	12.90	32.00	49.67	70.26	84.00	2.10	12.76	44.44	76.44	7.48	37.96
	rec	69.21	71.09	67.79	58.81	54.76	49.79	82.39	79.71	72.88	59.22	46.87	45.44
j	pre	8.34	21.47	35.75	48.53	58.03	65.25	2.81	10.19	27.81	47.25	32.16	60.31
	rec	75.42	55.45	35.06	26.60	20.94	17.71	51.14	29.40	21.12	16.94	93.29	33.42
k	pre	2.05	7.07	18.12	30.94	45.15	56.54	0.91	5.45	20.15	40.97	6.54	33.26
	rec	96.89	95.56	92.98	88.73	85.26	80.34	86.36	82.33	80.07	76.90	99.27	96.47
l	pre	2.08	7.22	18.76	32.57	47.81	61.87	1.03	6.24	23.19	47.14	6.35	32.10
	rec	95.48	94.67	93.41	90.61	87.56	85.26	95.45	91.34	89.40	85.82	93.56	90.30
m	pre	2.09	7.26	18.93	32.96	48.46	62.48	1.03	6.27	23.22	47.08	6.47	32.88
	rec	94.92	94.43	93.47	90.94	88.04	85.39	94.32	91.07	88.78	85.02	94.47	91.74
n	pre	1.89	6.56	17.14	29.52	41.01	48.89	0.05	5.93	21.81	41.48	5.50	26.86
	rec	99.15	98.55	97.76	94.07	86.05	77.18	100.00	99.46	96.28	86.51	92.75	86.56
o	pre	1.99	6.94	17.76	28.95	36.49	39.82	0.90	5.44	18.89	32.34	6.41	30.18
	rec	94.07	93.95	91.24	83.11	68.98	56.63	85.80	82.24	75.14	60.75	97.46	87.60
p	pre	1.52	5.32	14.02	25.05	38.00	50.25	0.76	4.77	18.10	38.29	4.70	24.25
	rec	100.00	100.00	100.00	99.83	99.71	99.19	100.00	100.00	99.93	99. 87	99.18	97.73
x	pre	1.99	6.98	18.34	32.74	49.37	64.21	0.99	6.25	23.57	49.21	6.16	31.59
	rec	100.00	100.00	99.82	99.57	98.85	96.72	100.00	100.00	99.29	97.93	99.18	97.14

down to 8-Common, C1 to 5-Common and C2 to 4-Common). Only files 'a', 'b', ('c', 'd', 'e',) 'g', 'h', 'i', 'l', 'n', and 'p' are relevant to C1, and 'f', 'j', 'k', 'm', and 'o' are relevant to C2, but we showed both C1 and C2 for all the files.

4.3 N-Common Evaluation Ignoring Weights

Table 15 shows the number of N-common Candidates appearing in each result file, for Ca-[10–5]-Candidates, C1-[7–4]-Candidates, and C2-[4–3]-Candidates. It can be observed, in general, that N-commons tend to appear more in bigger files. There are, however, some exceptions. For instance, the file 'o' takes a lower value with a smaller N, though 'o' is relatively large. Figures 14 and 15 illustrate the same information as in Table 15 ('x' is not included).

Table 16 shows the recall and precision of each file with respect to Ca-[10–5]-Candidates, C1-[7–4]-Candidates, and C2-[4–3]-Candidates. Here again, all the files are listed for C1 and C2. Figures 16 and 17 plot the same information as in Table 16.

Guessing from the very nature of N-common-Candidates, it would be expected that when N is large the smaller result files gain an advantage in that they score high in precision without losing recall, while as N decreases the precision of larger files become higher. The result shows this tendency rather straightforwardly. So it is not safe to conclude which result files performed better.

What is more interesting is the transition of the relative positions of the files in figures according to the changes of N. For instance, when N > 8 in Ca, the files 'g' and 'j' are very similar, but when N decreases they gradually become separate. Also, 'k', 'l', 'm', 'n', 'o', and 'p' show their individual characteristics when N becomes smaller.

4.4 N-Common Evaluation Considering Weights

We observed the precision of fixed intervals, using 1000 and 3000 as the intervals. The same method was used as in the evaluation based on Manual- and Index-Candidates. We used only two N-Common data, i.e. Ca-9-common and Ca-7-Common. These two were used because the number of Candidates became closer to Manual- and Index-Candidates.

The 10 files, i.e. 'a', 'b', 'f', 'h', 'j', 'l', 'm', 'n', 'o', and 'p', were examined with respect to the interval 1000, and 8 files, i.e. 'b', 'h', 'j', 'l', 'm', 'n', 'o', and 'p', were examined with respect to the interval 3000.

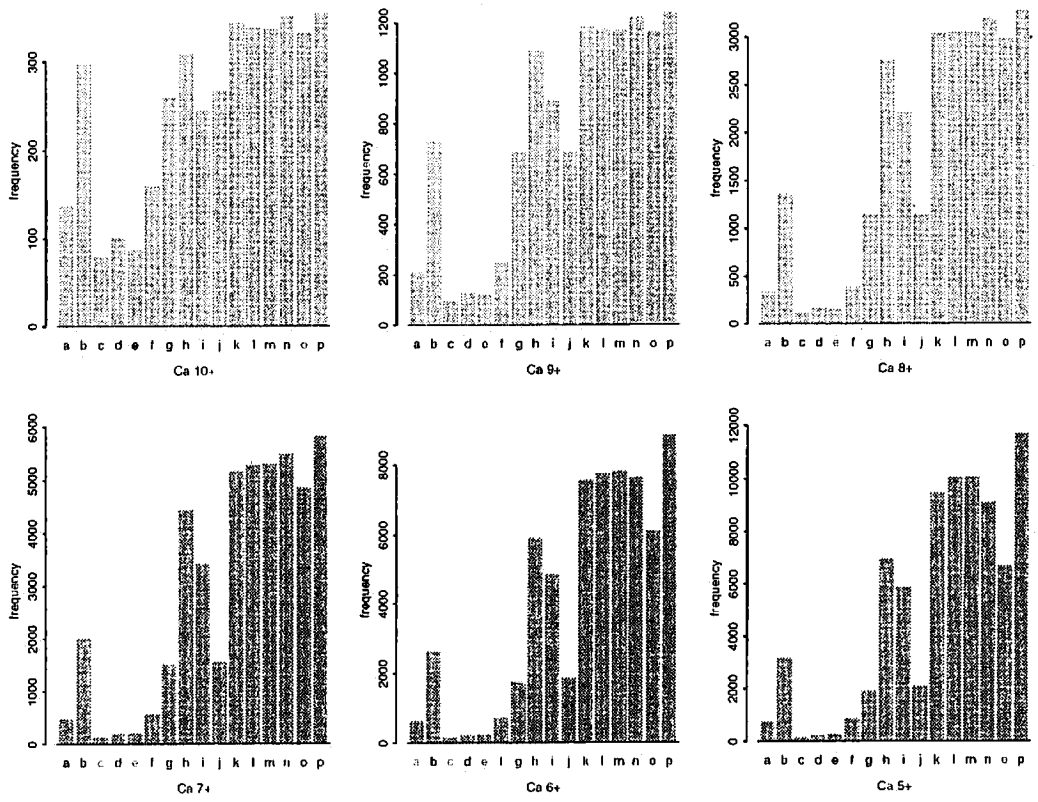


Figure 14 : Frequency of matched elements against Ca-N-common data

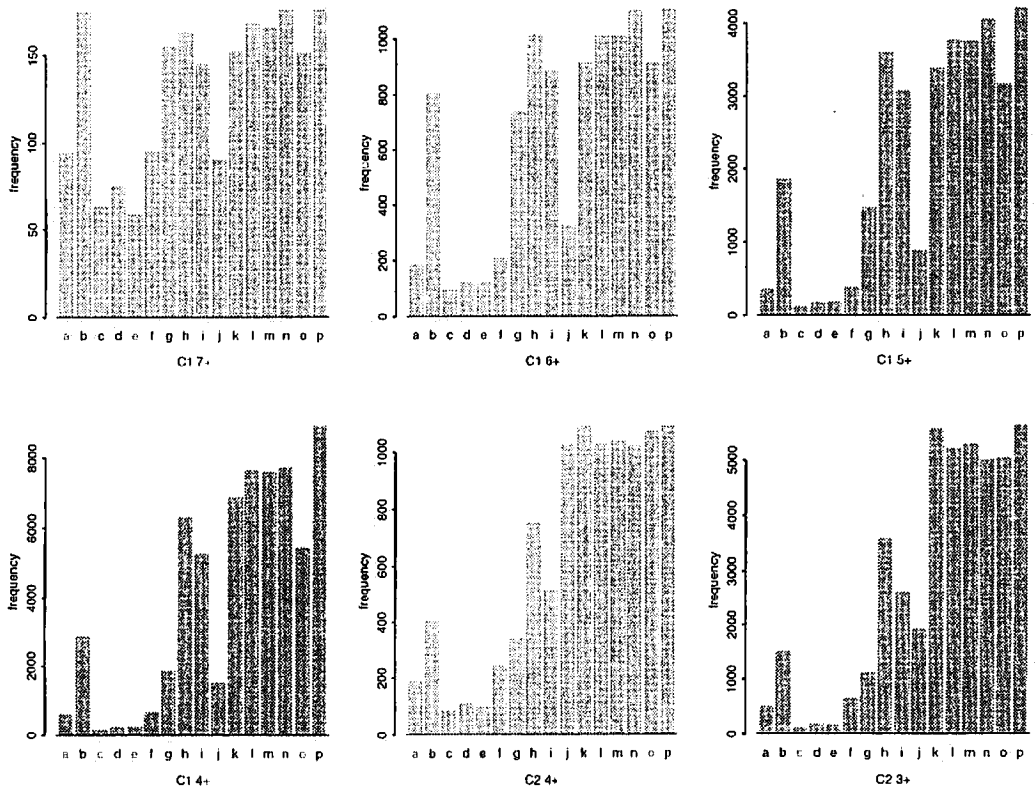


Figure 15 : Frequency of matched elements against C1- and C2-N-common data

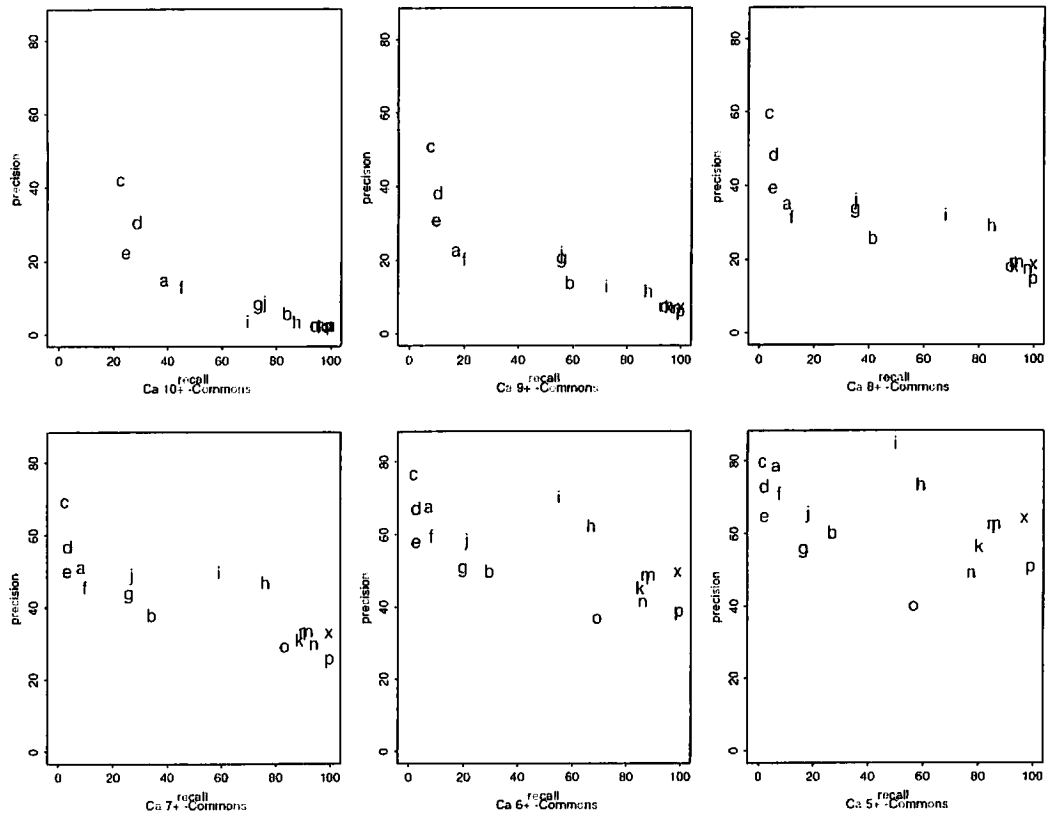


Figure 16 : Recall and precision against Ca–N-common

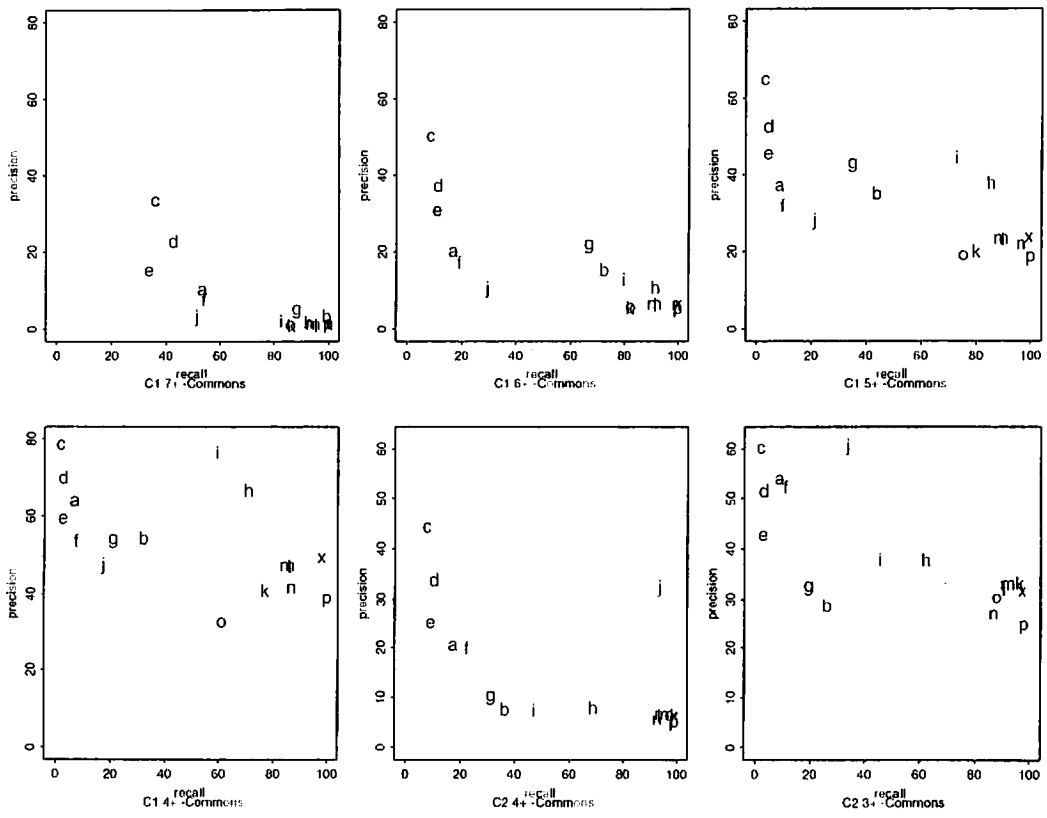


Figure 17 : Recall and precision against C1- and C2-N-common

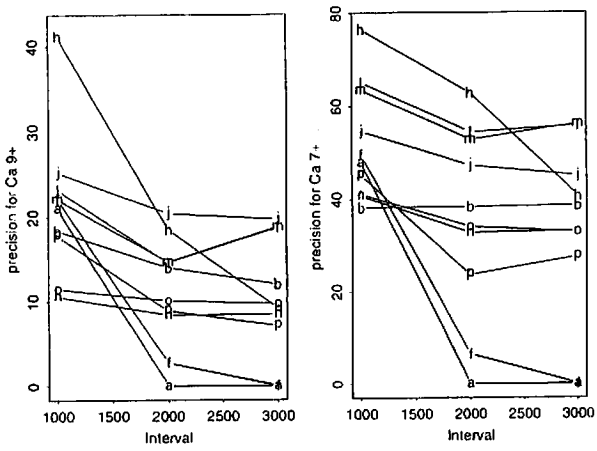


Figure 18 : Up to 3000 in intervals of 1000 based on Ca–N-common

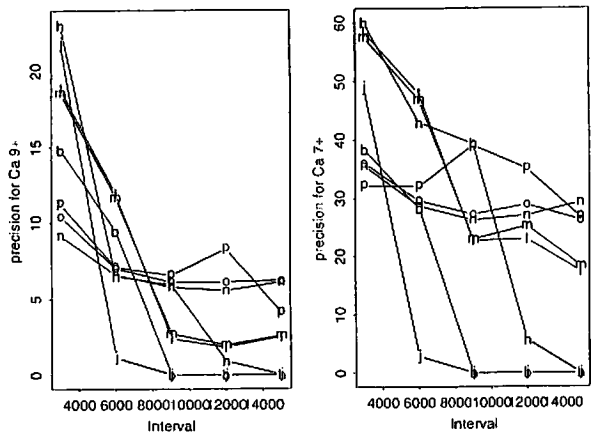


Figure 19 : Up to 15000 in intervals of 3000 based on Ca–N-common

The values for precision are based on each interval, not on cumulative intervals. Figures 18 and 19 show the results for intervals 1000 and 3000, respectively. Such files as 'n' and 'o', for instance, which show stable values for all the intervals, are considered to extract different types of candidate, while the files that show a steep downward curve from left to right, such as 'h', are considered to be very neutral among the result files.

5 On the Similarity of Result Files

We examine here the similarity of result files (unfortunately, we could not include 'x' here). An idea that immediately springs to mind is to apply some clustering algorithms to the result files. Here, however, we focus on the similarity between file pairs, because the sizes of the data are very different.

5.1 Ignoring Weights

Table 17 shows the number of common elements among the top 1000 term candidates in the files 'a', 'b', 'f', 'g', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', and 'p'. As the data in the files 'g', 'i', and 'k' are not weighted, we used the average value of the 500 random extraction of 1000 elements. The average value for each file is calculated excluding the files that belong to the same team.

Table 18 shows the number of common elements among the top 3000 term candidates in the 11 files, i.e. 'b', 'g', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', and 'p'. For 'g', 'i', and 'k', we used the average of 500-time random extraction.

It can be observed from here that 'h' has a greater number of common elements on average, while 'i', 'k', 'n', and 'o' have a smaller number of common elements.

5.2 Considering Weights

Table 19 shows the pairwise values of Spearman's rank-coefficient. The last row, i.e. 'mean', indicates the average excluding the result files from the same team. The file 'h' takes a notably high average value, followed by 'l' and 'm'. Interesting file pair correlation is observed between 'b' on the one hand and 'n' and 'o' on the other, as well as between 'p' and 'n' and 'o'. They show negative correlations. 'a', 'n', and 'p' show on average smaller values of the coefficient.

The standard Spearman's rank-coefficient ρ requires that two data have the same number of items. So we cannot apply this to compare files of different sizes. Here, therefore, we applied the rank-coefficient to the common items in each file pair. The scatterplot of the pairs of files 'a', 'b', 'f', 'h', 'j', 'l', 'm', 'n', 'o', and 'p' is shown in Figure 20.

6 Conclusions

We have so far observed the characteristics of result files from several points of view. There are still many others remaining, such as N-Common observation by length, etc. However, the basic information necessary for a constructive discussion concerning the different methods for term recognition was provided in this report. We used such terms as 'precision' or 'recall', which imply whether something is good or not, because it was convenient from the point of view of characterising the result files. However, we do not think that 'precision' or 'recall' show the superiority or inferiority of the result files at all. There are a few reasons for this:

- We do not think that the common understanding of the very concepts of 'terms' and 'terminology' are at such a level as to be able to evaluate different results from a single point of view on the basis of 'recall' and/or 'precision'. Our Manual-Candidates and Index-Candidates are reflections of *a possible* conceptualisation of 'terms' and 'terminology', but we do not

Table 17 : Number of common elements among the top 1000 elements

	a	b	f	g	h	i	j	k	l	m	n	o	p
a	1000	214	777	45	109	51	65	43	38	35	41	38	202
b	214	1000	204	51	137	50	31	42	69	65	45	35	144
f	777	204	1000	51	117	57	60	46	46	43	41	39	222
g	45	51	51	(1000)	113	122	64	96	66	64	54	51	61
h	109	137	117	113	1000	53	183	45	191	183	142	151	111
i	51	50	57	122	53	(1000)	49	24	73	73	28	18	62
j	65	31	60	64	183	49	1000	37	205	210	70	59	59
k	43	42	46	96	45	24	37	(1000)	47	45	33	29	47
l	38	69	46	66	191	73	205	47	1000	942	42	34	77
m	35	65	43	64	183	73	210	45	942	1000	39	32	75
n	41	45	41	54	142	28	70	33	42	39	1000	598	65
o	38	35	39	51	151	18	59	29	34	32	598	1000	63
p	202	144	222	61	111	62	59	47	77	75	65	63	1000
mean	66.7	66.9	73.1	69.8	127.9	55.5	95.6	44.5	80.7	78.5	54.5	49.9	99.0

Table 18 : Number of common elements among the top 3000 elements

	b	g	h	i	j	k	l	m	n	o	p
b	3000	368	577	437	297	365	577	577	322	307	631
g	368	3000	595	369	443	292	505	497	470	434	414
h	577	595	3000	478	682	372	972	949	527	566	457
i	437	369	478	3000	410	169	597	600	262	187	495
j	297	443	682	410	3000	319	762	761	390	366	519
k	365	292	372	169	319	3000	401	394	254	224	372
l	577	505	972	597	762	401	3000	2797	413	370	571
m	577	497	949	600	761	394	2797	3000	415	365	598
n	322	470	527	262	390	254	413	415	3000	1733	236
o	307	434	566	187	366	224	370	365	1733	3000	207
p	631	414	457	495	519	372	571	598	236	207	3000
mean	445.8	438.7	617.5	399.3	504.3	316.2	574.2	572.9	365.4	336.2	450.0

Table 19 : Spearman's rank-coefficient among shared elements

	a	b	f	h	j	l	m	n	o	p
a	1	0.082	0.905	0.022	0.017	0.019	0.013	0.087	0.103	0.041
b	0.082	1	0.088	0.101	0.012	0.165	0.171	-0.031	-0.034	0.133
f	0.905	0.088	1	0.083	-0.059	0.095	0.098	0.078	0.134	0.090
h	0.022	0.101	0.083	1	0.251	0.325	0.330	0.274	0.249	0.210
j	0.017	0.012	-0.059	0.251	1	0.228	0.236	0.021	0.014	0.071
l	0.019	0.165	0.095	0.325	0.228	1	0.987	0.083	0.105	0.081
m	0.013	0.171	0.098	0.330	0.236	0.987	1	0.089	0.075	0.062
n	0.087	-0.031	0.078	0.274	0.021	0.083	0.089	1	0.915	-0.092
o	0.103	-0.034	0.134	0.249	0.014	0.105	0.075	0.915	1	-0.009
p	0.041	0.133	0.090	0.210	0.071	0.081	0.062	-0.092	-0.009	1
mean	0.043	0.087	0.074	0.215	0.088	0.138	0.134	0.064	0.080	0.065

claim that we can give any special status to it. Rather, it is intended to be used for stimulating the further discussion.

- As 'terms' are first and foremost consolidated as part of 'terminology' (Kageura, 1999), we believe that the internal uniformity or consistency of the candidates among the extracted set should be an essential criterion for evaluation, which we could not pursue here. It is not possible to know by evaluating individual terms whether we are doing keyword extraction vis-a-vis documents or term extraction.

This paper is intended to give comparative analyses of result files submitted by participants, in order to give a concrete starting point for constructive discussions in automatic term recognition. We hope this paper contributed to that purpose.

References

Shapiro, S. (1987) *Encyclopedia of Artificial Intelligence*. New York: John Wiley. [Ohsuga, S. (trans. ed.) *Jinko Tinou Daijiten*. Tokyo: Maruzen . 1991]

Kageura, K. (1999) "On *Quid Juris* of the theoretical study of terminology, and a sketch of a possible framework for the theoretical study of term formation and terminological growth (or when quality meets quantity)," *Primero Seminario de Terminologia Teorica*, 28–29 January, Barcelona: University of Pompeu Fabra. (available from http://www.rd.nacsis.ac.jp/~kyo/alist.html)

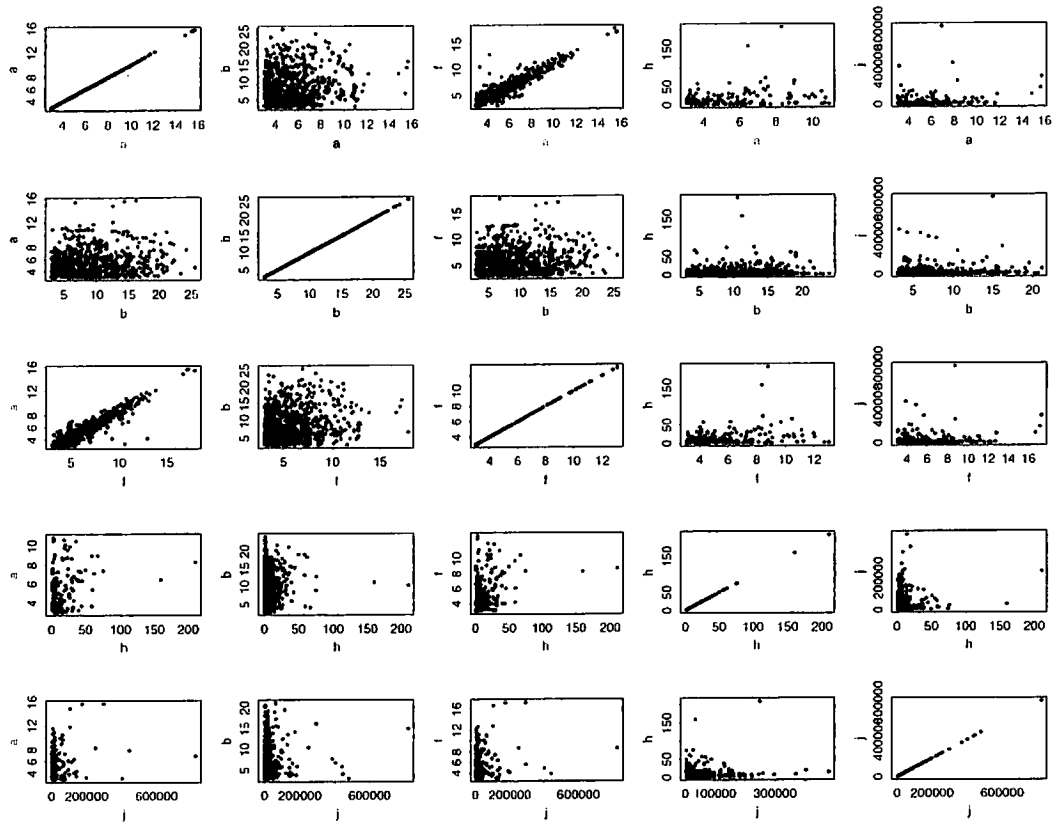


Figure 20(a) : Correlation scatterplot (a,b,f,h,j × a,b,f,h,j)

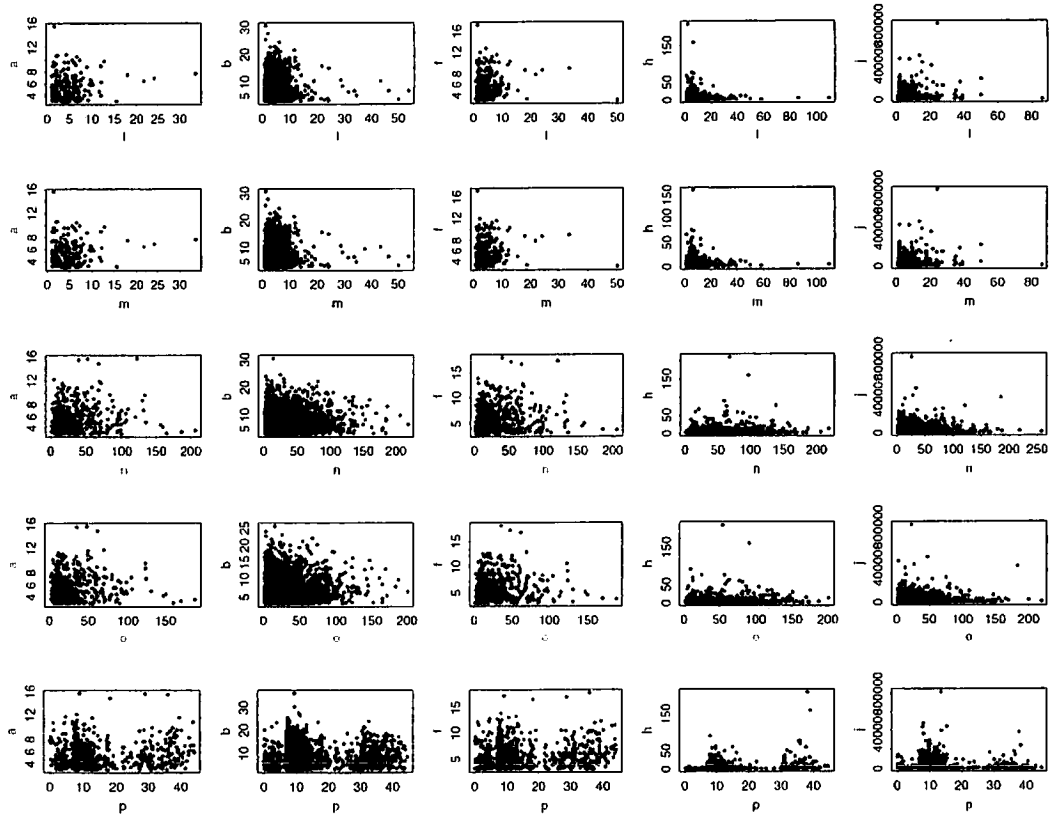


Figure 20(b) : Correlation scatterplot (a,b,f,h,j × l,m,o,p,q)

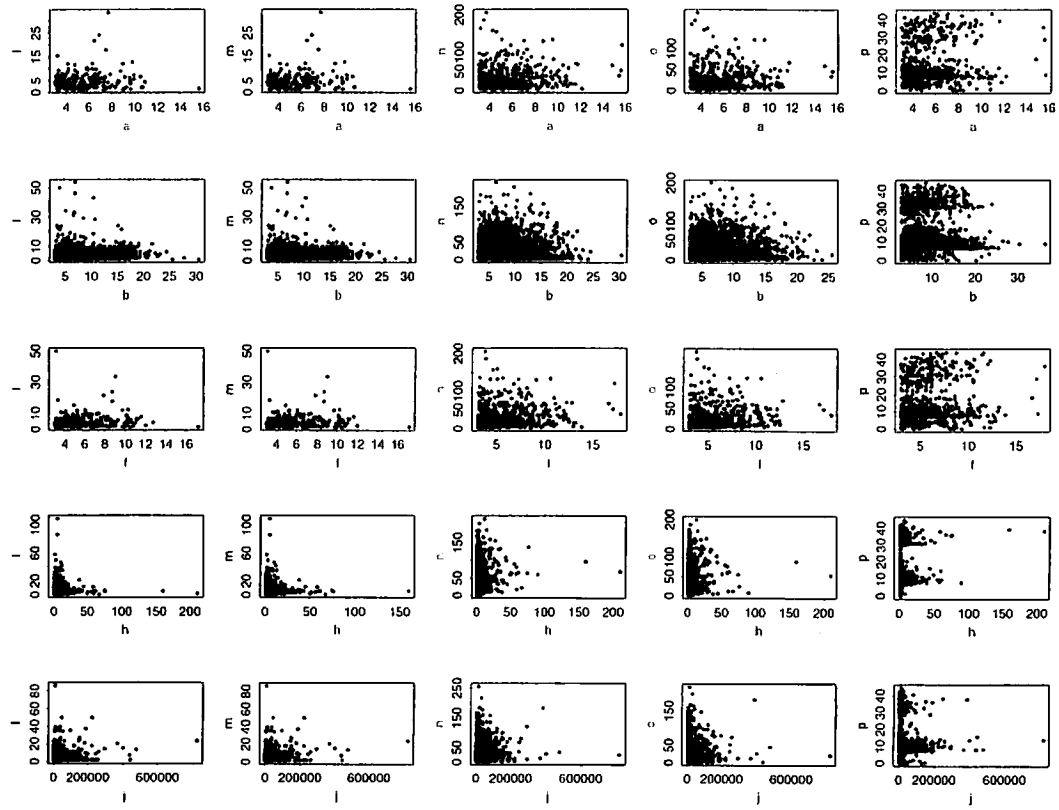


Figure 20(c) : Correlation scatterplot (l,m,o,p,q × a,b,f,h,j)

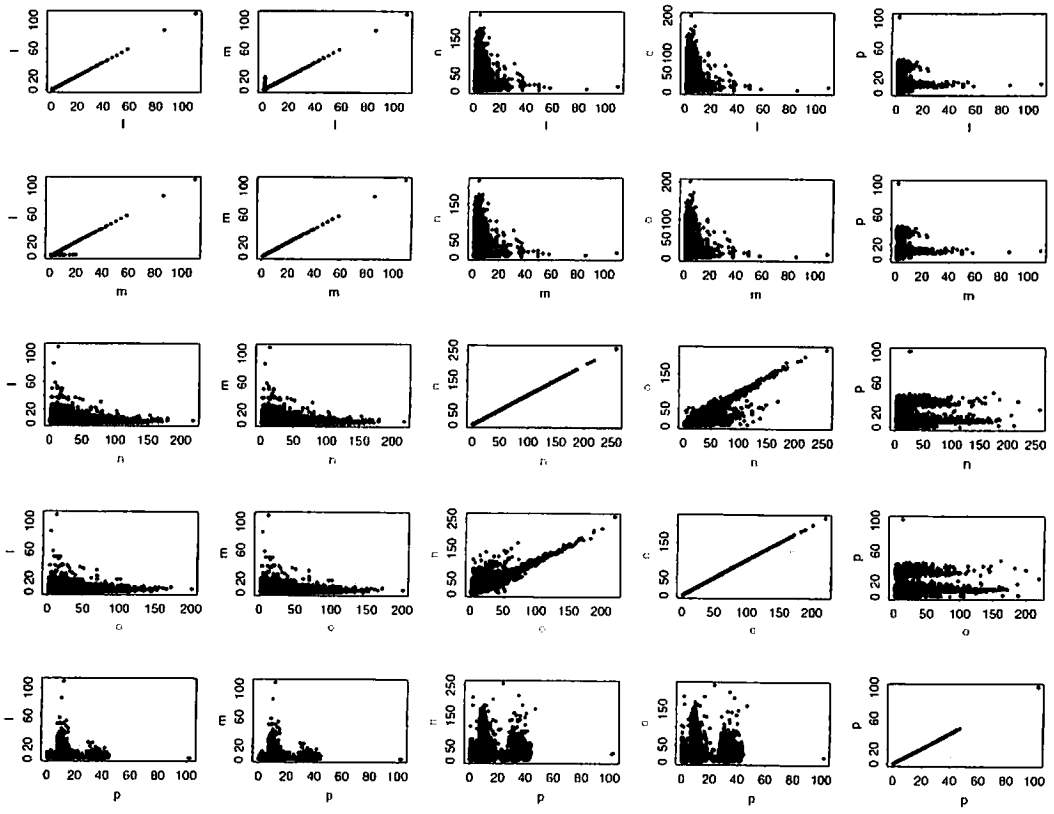


Figure 20(d) : Correlation scatterplot (l,m,o,p,q × l,m,o,p,q)