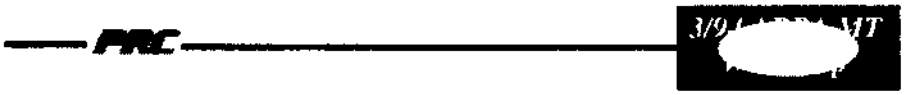
RESULTS

1Q94 ARPA EVALUATION OF MACHINE TRANSLATION

March 17, 1994



Part 3 - System Test Results

- Computation Methods
- Comprehension, Fluency, Adequacy
  - Comparison with 2Q93
  - FAMT systems by language

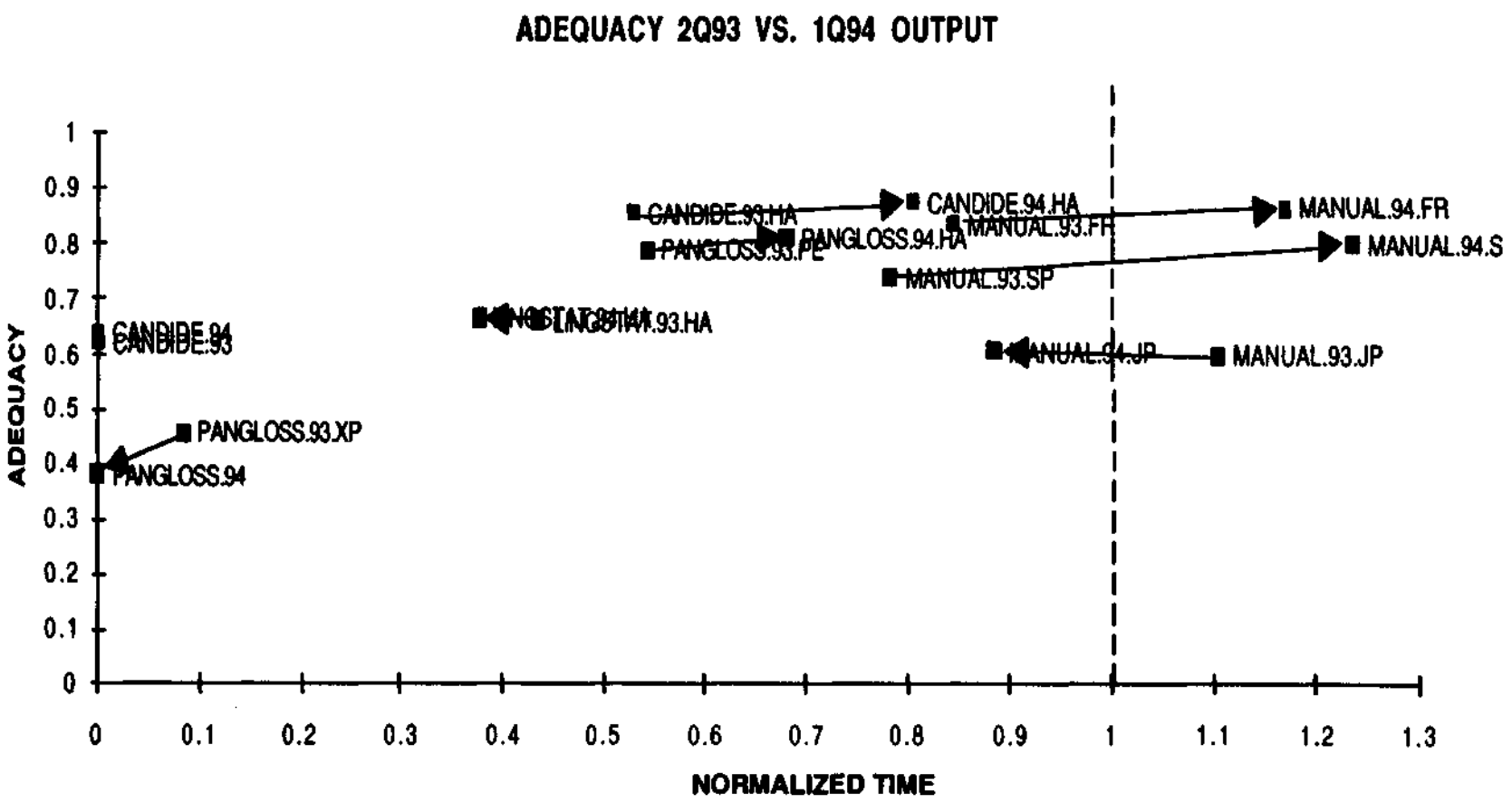
Time Computation

- Time axis = Translation time /AVG. (manual 93 time + manual 94 time)
- Mean of operator time, manual time for each system
- Presumes indirect comparability

Comprehension, Fluency and Adequacy Computation

- Passage Score between 0 and 1
  - $\text{Comprehension}_{(P)}$ = #Correct/6
  - $\text{Fluency}_{(P)} = (\sum((\text{Judgment point} - 1)/(5-1)))/\#\text{sent. in passage}$
  - $\text{Adequacy}_{(P)} = (\sum((\text{Judgment point} - 1)/(5-1)))/\#\text{frags. in passage}$
- System Score:
  - Mean and STD DEV calculated for passage score over 20 passages
- STD DEV calculated over all system scores
- F-Ratio as ((VAR of System Means / mean of System VARS)

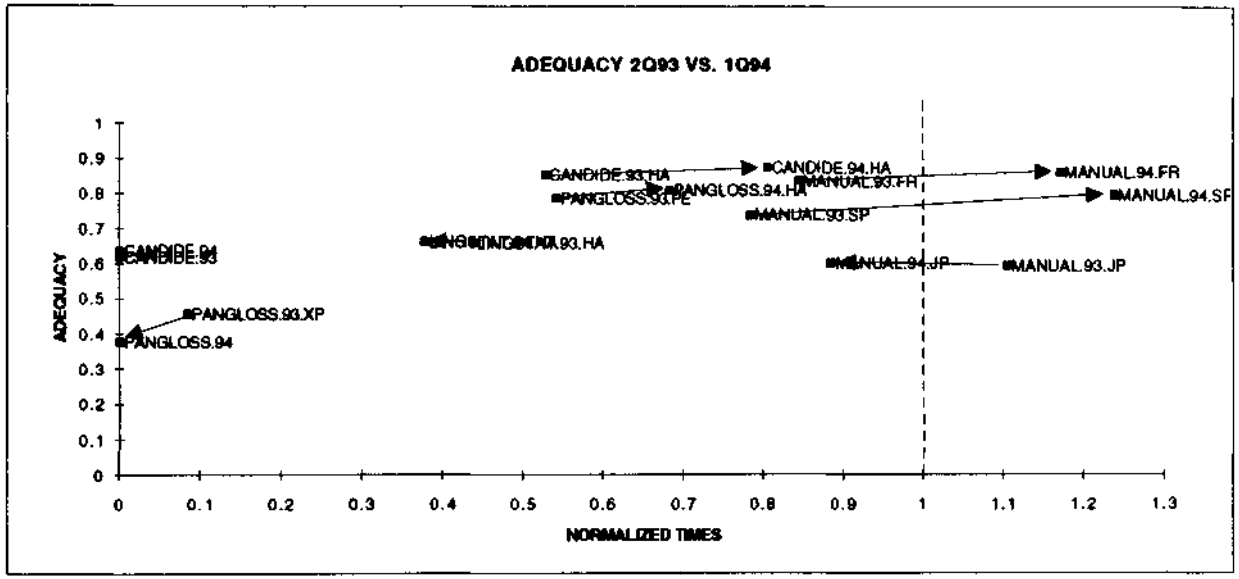
Adequacy Comparison with 2Q93



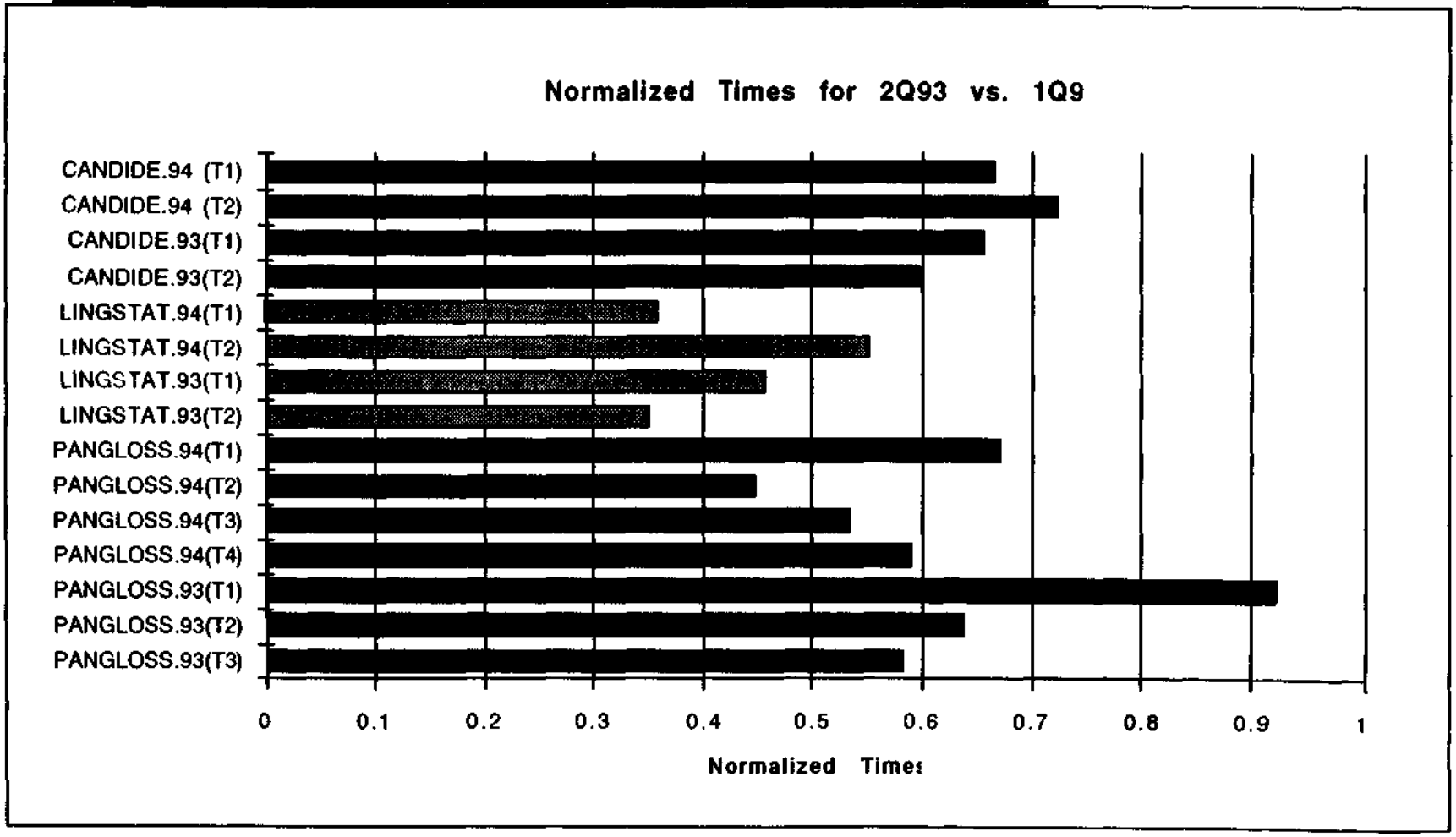
ADEQUACY 2Q93 VS. 1Q94 OUTPUT - 1Q94 MT EVALUATION

	CANDIDE.94	CANDIDE.94.HA	PANGLOSS.94	PANGLOSS.94.HA	LINGSTAT.94.HA	MANUAL.94.FR	MANUAL.94.JP	MANUAL.94.SP	MEAN	STD DEV	VARIANCE
(Pass.Scores/5)/ Num of Sentences	0.638	0.875	0.377	0.810	0.663	0.863	0.604	0.797	0.703	0.157	0.025
											Total F Ratio
Var of 20 pts	0.031	0.008	0.032	0.019	0.025	0.007	0.038	0.020	0.023		4.892
std dev of 20 pts	0.175	0.088	0.180	0.139	0.158	0.086	0.195	0.142	0.145		
Norm Time	0.000	0.805	0.000	0.682	0.377	1.171	0.885	1.237	0.645	0.449	AVG STD DEV
Var of 20 pts	0.000	0.101	0.000	0.037	0.005	0.212	0.105	0.119	0.072		0.033
std dev of 20 pts	0.000	0.318	0.000	0.194	0.072	0.460	0.325	0.336	0.213		

	CANDIDE.93	CANDIDE.93.HA	PANGLOSS.93.XP	PANGLOSS.93.PE	LINGSTAT.93.HA	MANUAL.93.FR	MANUAL.93.JP	MANUAL.93.SP	MEAN	STD DEV	VARIANCE
(Pass.Scores/5)/ Num of Sentences	0.619	0.854	0.456	0.787	0.660	0.836	0.595	0.739	0.693	0.127	0.016
											Total F Ratio
Var of 22 pts	0.017	0.019	0.031	0.015	0.021	0.011	0.034	0.017	0.021		3.664
std dev of 22 pts	0.131	0.140	0.175	0.122	0.146	0.107	0.184	0.132	0.142		
Norm Time	0.000	0.528	0.084	0.542	0.437	0.845	1.105	0.784	0.540	0.350	AVG STD DEV
Var of 22 times	0.000	0.028	0.018	0.026	0.003	0.058	0.067	0.066			0.050
std dev of 22 times	0.000	0.168	0.133	0.162	0.050	0.242	0.259	0.256			



HAMT / Manual by Translator



1994 MT Evaluation

CANDIDE.94	
T1	T2
567/849	325/448
0.668	0.725

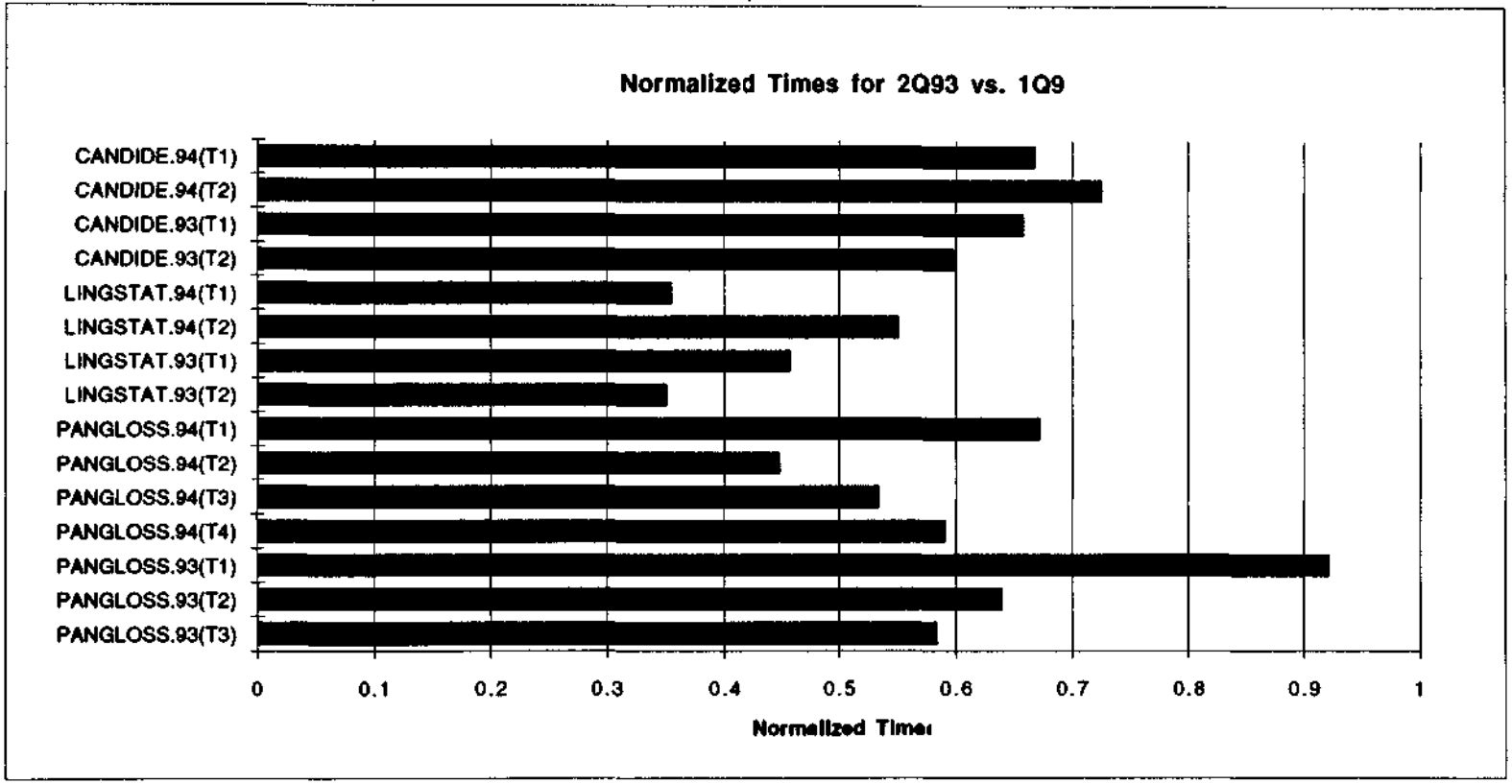
CANDIDE.93	
T1	T2
299/454	344/575
0.659	0.598

LINGSTAT.94	
T1	T2
969/2728	849/1539
0.355	0.552

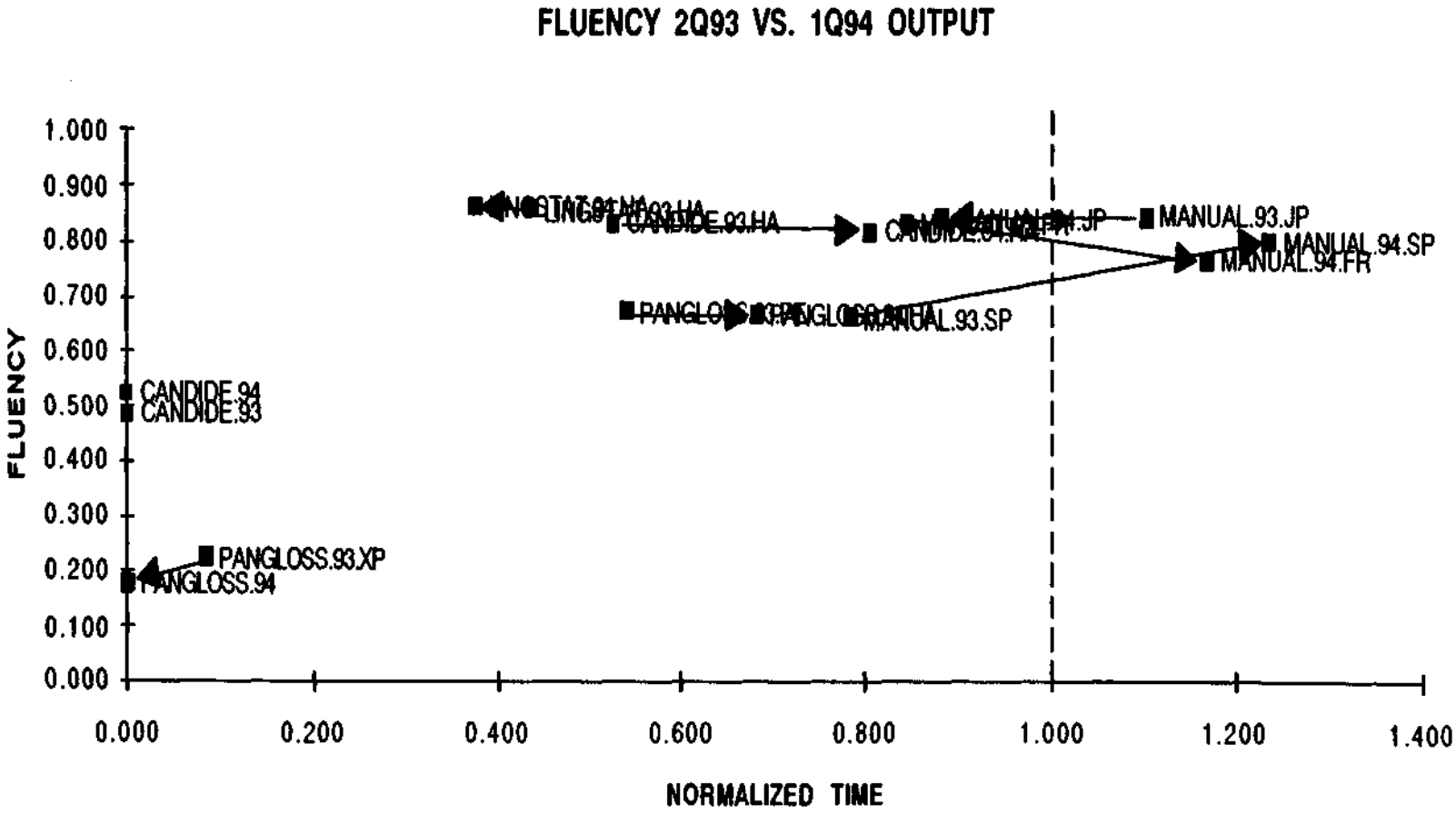
LINGSTAT.93	
T1	T2
1114/2432	1203/3428
0.458	0.351

PANGLOSS.94			
T1	T2	T3	T4
277/412	249/555	171/320	246/416
0.672	0.449	0.534	0.591

PANGLOSS.93		
T1	T2	T3
294/319	217/339	309/529
0.922	0.640	0.584



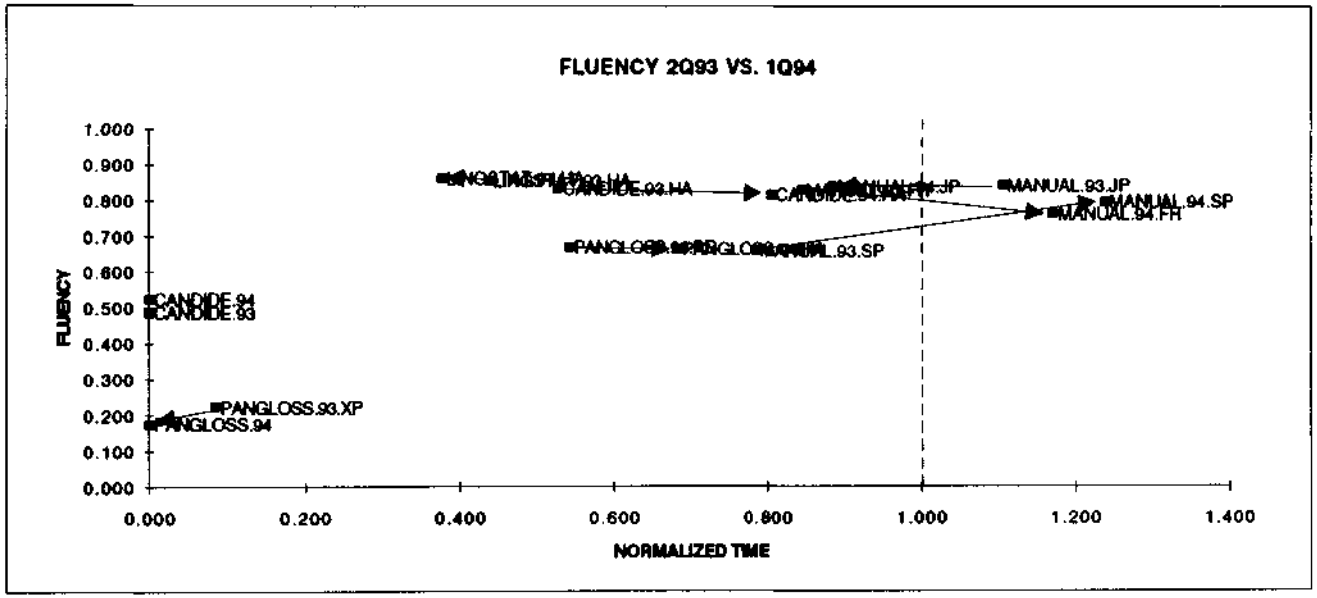
Fluency Comparison with 2Q93



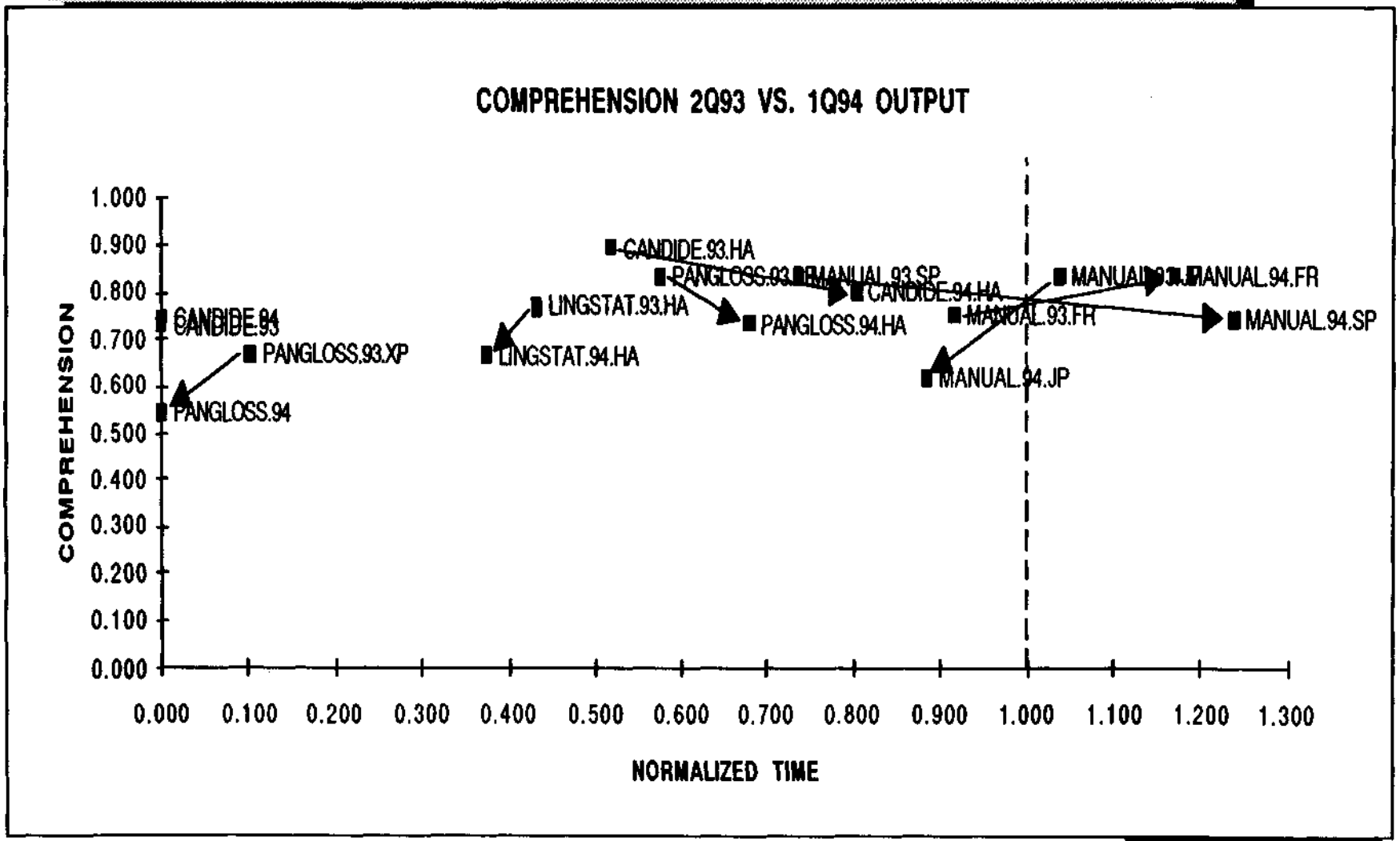
FLUENCY 2Q93 VS. 1Q94 OUTPUT - 1Q94 MT EVALUATION

	CANDIDE.94	CANDIDE.94.HA	PANGLOSS.94	PANGLOSS.94.HA	LINGSTAT.94.HA	MANUAL.94.FR	MANUAL.94.JP	MANUAL.94.SP	MEAN	STD DEV	VARIANCE
(Pass.Scores/5)/ Num of Sentences	0.524	0.817	0.176	0.665	0.662	0.764	0.841	0.797	0.681	0.217	0.047
											Total F Ratio
Var of 20 pts	0.051	0.021	0.019	0.039	0.016	0.032	0.018	0.029	0.028		7.466
std dev of 20 pts	0.225	0.145	0.139	0.198	0.127	0.178	0.135	0.171	0.165		
Norm Time	0.000	0.805	0.000	0.682	0.377	1.171	0.885	1.237	0.645	0.449	AVG STD DEV
Var of 20 pts	0.000	0.101	0.000	0.037	0.005	0.212	0.105	0.113	0.072		0.037
std dev of 20 pts	0.000	0.318	0.000	0.194	0.072	0.460	0.325	0.336	0.213		

	CANDIDE.93	CANDIDE.93.HA	PANGLOSS.93.XP	PANGLOSS.93.PE	LINGSTAT.93.HA	MANUAL.93.FR	MANUAL.93.JP	MANUAL.93.SP	MEAN	STD DEV	VARIANCE
(Pass.Scores/5)/ Num of Sentences	0.486	0.833	0.223	0.670	0.856	0.830	0.843	0.660	0.675	0.209	0.044
											Total F Ratio
Var of 22 pts	0.027	0.014	0.016	0.015	0.009	0.016	0.013	0.020	0.016		12.584
std dev of 22 pts	0.165	0.118	0.128	0.124	0.093	0.127	0.114	0.142	0.126		
Norm Time	0.000	0.528	0.084	0.542	0.437	0.845	1.105	0.784	0.540	0.350	AVG STD DEV
Var of 22 times	0.000	0.028	0.018	0.026	0.003	0.058	0.067	0.066			0.045
std dev of 22 times	0.000	0.168	0.133	0.162	0.050	0.242	0.259	0.256			



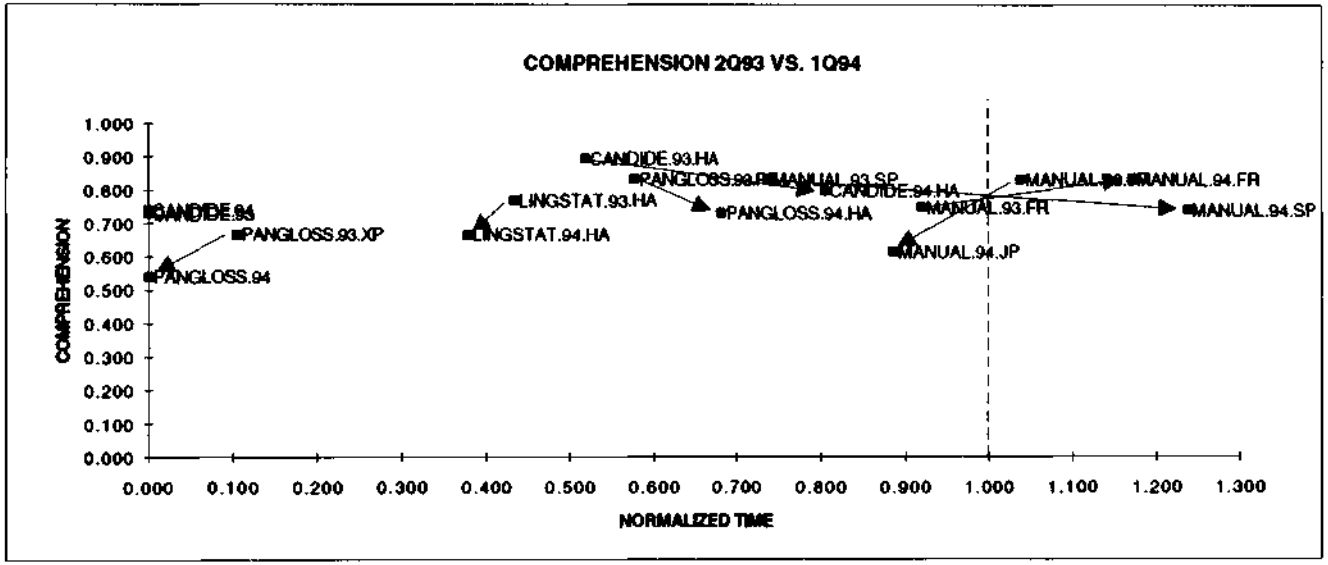
Comprehension Comparison with 2Q93



COMPREHENSION 2Q93 VS. 1Q94 OUTPUT - 1Q94 MT EVALUATION

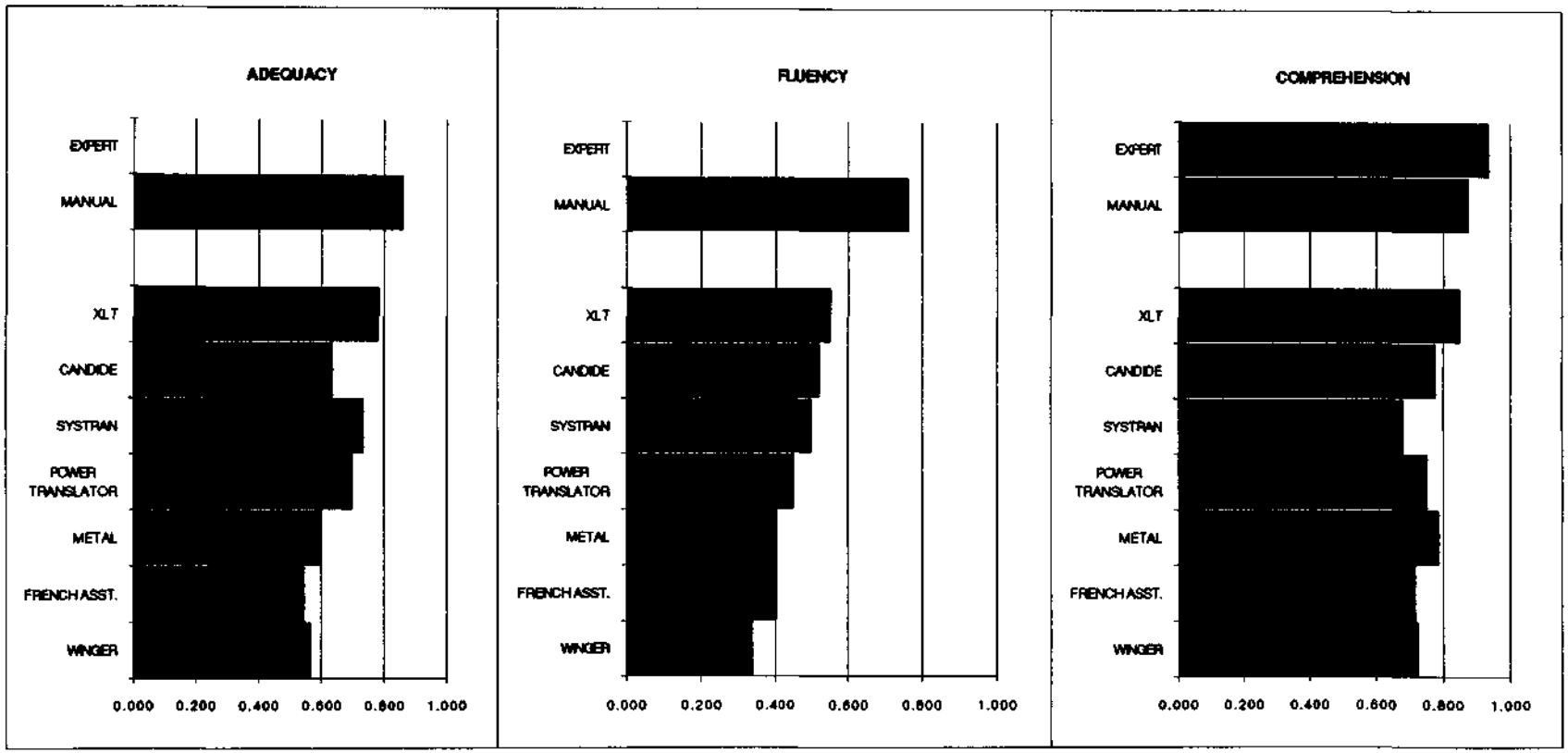
	CANDIDE.94	CANDIDE.94.HA	PANGLOSS.94	PANGLOSS.94.HA	LINGSTAT.94.HA	MANUAL.94.FR	MANUAL.94.JP	MANUAL.94.SP	MEAN	STD DEV	VARIANCE
(Pass.Scores/6)/ Num of Questions	0.742	0.800	0.542	0.733	0.667	0.833	0.617	0.742	0.709	0.090	0.008
											Total F Ratio
Var of 20 pts	0.032	0.066	0.055	0.048	0.053	0.042	0.059	0.029	0.048		0.756
std dev of 20 pts	0.179	0.256	0.235	0.220	0.230	0.204	0.242	0.171	0.217		
Norm Time	0.000	0.805	0.000	0.682	0.377	1.171	0.885	1.237	0.645	0.449	AVG STD DEV
Var of 20 pts	0.000	0.101	0.000	0.037	0.005	0.212	0.105	0.113	0.072		0.049
std dev of 20 pts	0.000	0.318	0.000	0.194	0.072	0.460	0.325	0.336	0.213		

	CANDIDE.93	CANDIDE.93.HA	PANGLOSS.93.XP	PANGLOSS.93.PE	LINGSTAT.93.HA	MANUAL.93.FR	MANUAL.93.JP	MANUAL.93.SP	MEAN	STD DEV	VARIANCE
(Pass.Scores/6)/ Num of Questions	0.729	0.896	0.667	0.833	0.771	0.750	0.833	0.833	0.789	0.069	0.005
											Total F Ratio
Var of 8 pts	0.023	0.023	0.063	0.016	0.031	0.048	0.040	0.040	0.036		0.375
std dev of 8 pts	0.153	0.153	0.252	0.126	0.177	0.218	0.199	0.199	0.185		
Norm Time	0.000	0.519	0.104	0.577	0.433	0.919	1.038	0.738	0.541	0.340	AVG STD DEV
Var of 8 pts	0.000	0.049	0.103	0.025	0.000	0.092	0.074	0.053	0.049		0.065
std dev of 8 pts	0.000	0.220	0.011	0.157	0.017	0.304	0.273	0.230	0.151		



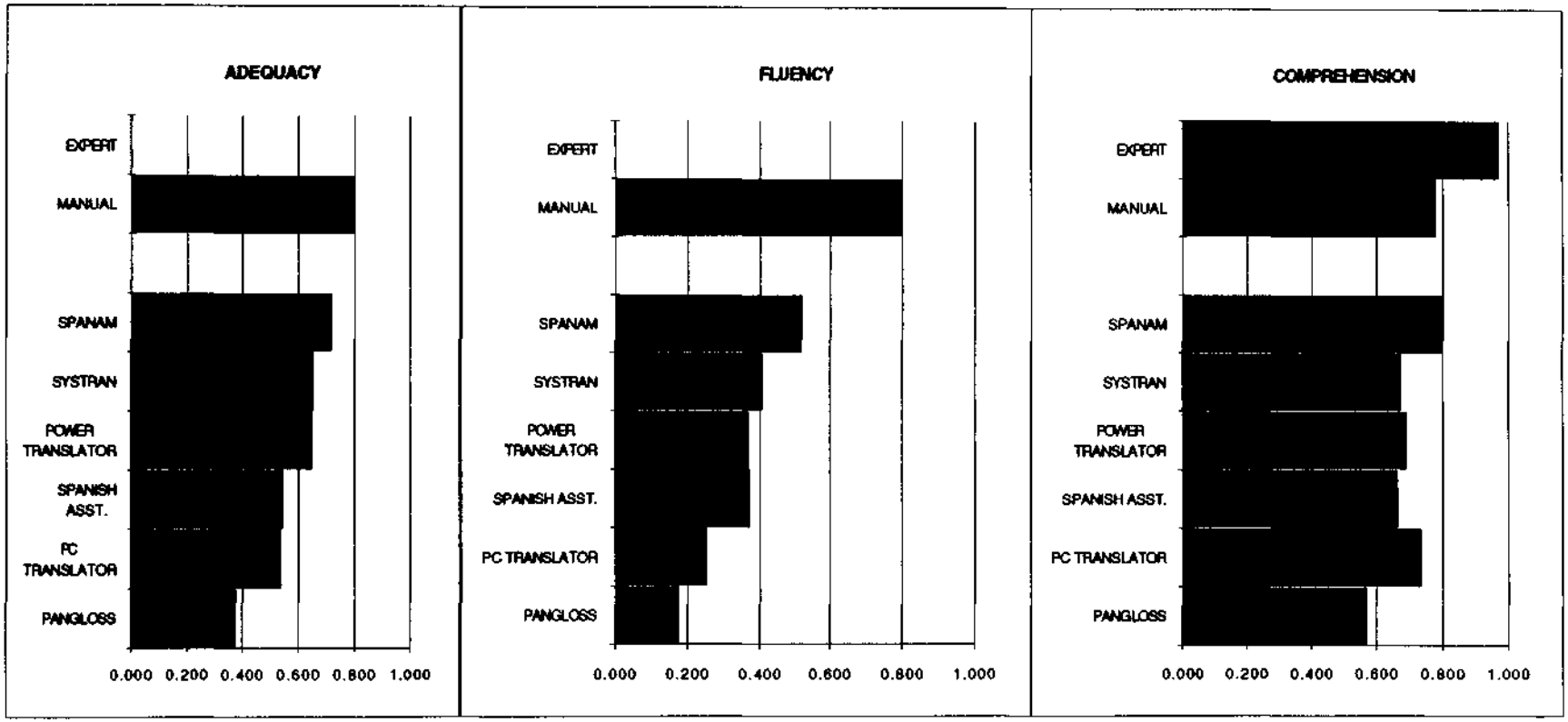
FAMT EVALUATION FOR FRENCH, 1Q94

	CANDIDE	SYSTRAN	MANUAL	EXPERT	POWER TRANSLATOR	FRENCH ASST.	METAL	XLT	WINGER	MEAN	STD DEV	VARIANCE	F-RATIO	AVG STD DEV
COMPREHENSION	0.781	0.684	0.877	0.939	0.754	0.719	0.789	0.851	0.728	0.791	0.083	0.007	0.670	0.047
VARIANCE	0.034	0.064	0.044	0.021	0.047	0.061	0.054	0.030	0.056	0.046				
STD DEVIATION	0.183	0.253	0.209	0.146	0.217	0.247	0.233	0.173	0.237	0.211				
FLUENCY	0.524	0.501	0.764		0.454	0.405	0.406	0.554	0.339	0.493	0.130	0.017	F-RATIO	AVG STD DEV
VARIANCE	0.051	0.025	0.032		0.020	0.020	0.031	0.018	0.028	0.026			2.689	0.037
STD DEVIATION	0.225	0.159	0.178		0.140	0.143	0.177	0.133	0.169	0.165				
ADEQUACY	0.638	0.736	0.863		0.700	0.548	0.598	0.786	0.566	0.679	0.111	0.012	F-RATIO	AVG STD DEV
VARIANCE	0.031	0.020	0.007		0.016	0.047	0.037	0.019	0.036	0.027			2.087	0.035
STD DEVIATION	0.175	0.140	0.086		0.127	0.216	0.193	0.138	0.191	0.158				

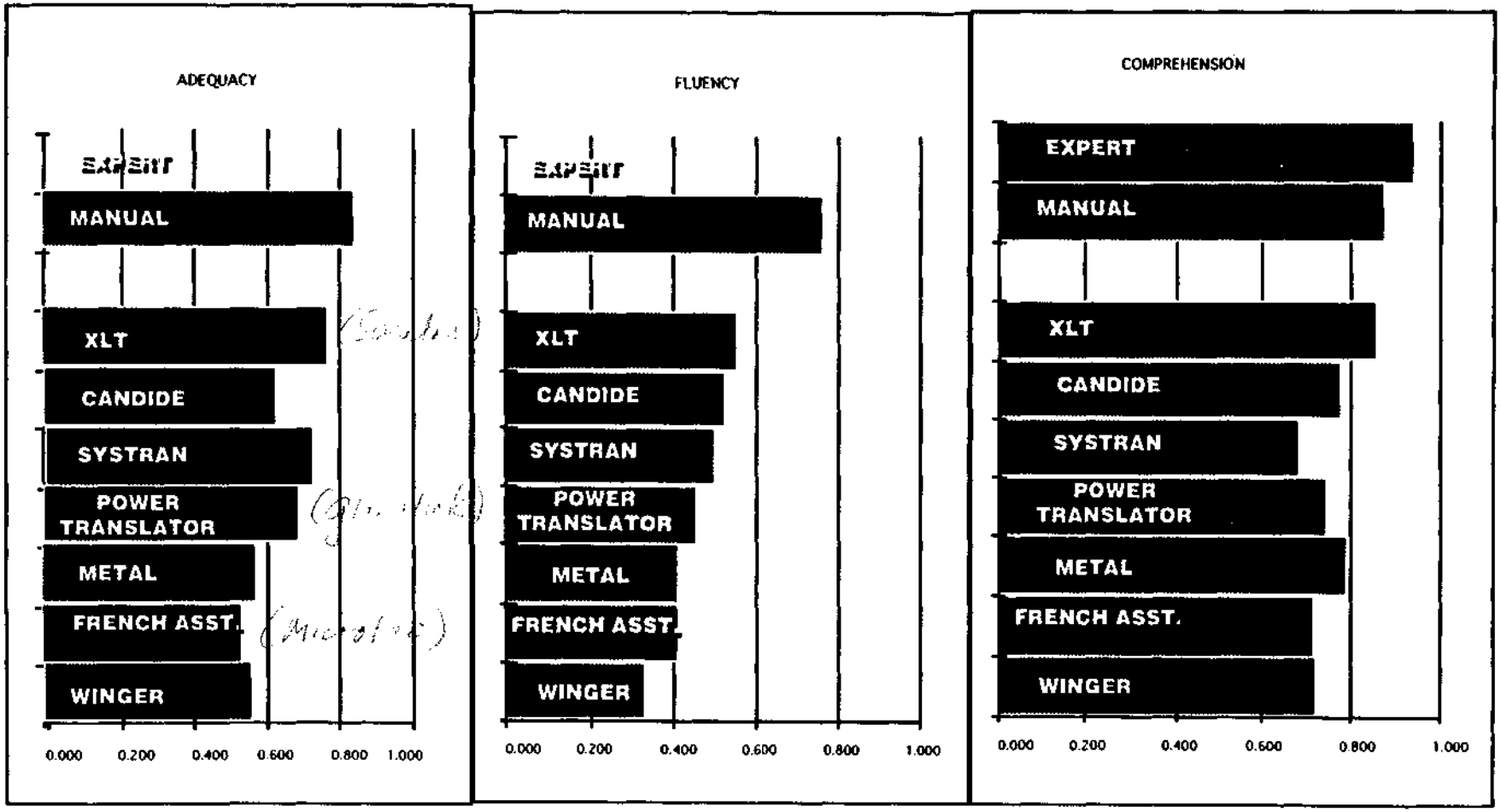


FAMT EVALUATION FOR SPANISH, 1Q94

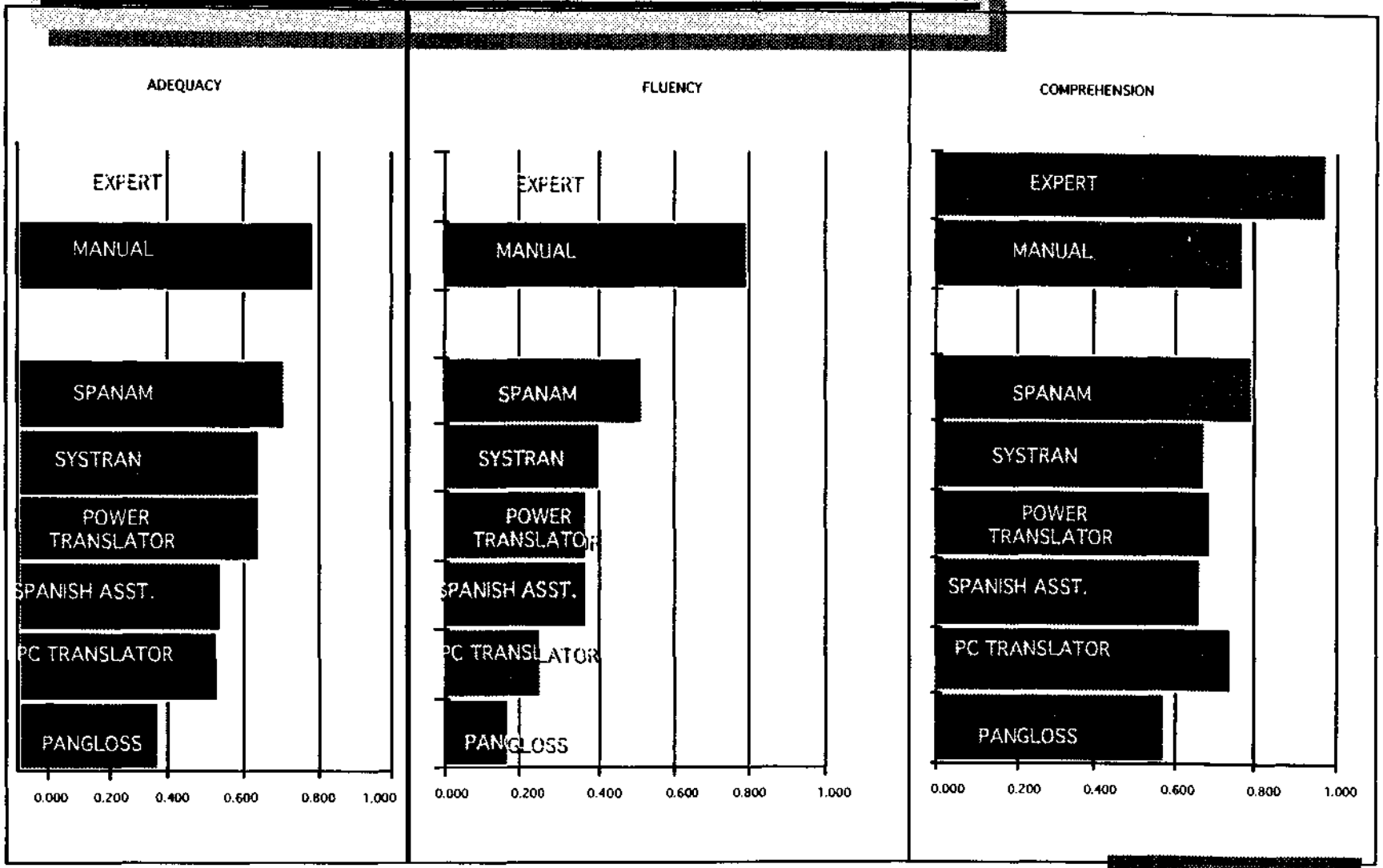
	PANGLOSS	SPANAM	MANUAL	EXPERT	POWER TRANSLATOR	SPANISH ASST.	SYSTRAN	PC TRANSLATOR	MEAN	STD DEV	VARIANCE	F-RATIO	AVG STD DEV
COMPREHENSION	0.570	0.798	0.781	0.974	0.693	0.667	0.675	0.737	0.737	0.120	0.014	1.240	0.049
VARIANCE	0.058	0.045	0.031	0.010	0.063	0.069	0.071	0.066	0.052				
STD DEV	0.241	0.213	0.175	0.101	0.251	0.263	0.266	0.257	0.221				
FLUENCY	0.176	0.520	0.797		0.369	0.373	0.408	0.252	0.414	0.202	0.041	F-RATIO	AVG STD DEV
VARIANCE	0.019	0.032	0.029		0.024	0.051	0.019	0.022	0.028			6.458	0.037
STD DEV	0.139	0.180	0.171		0.155	0.226	0.139	0.149	0.166				
ADEQUACY	0.377	0.719	0.797		0.650	0.546	0.656	0.540	0.612	0.138	0.019	F-RATIO	AVG STD DEV
VARIANCE	0.032	0.021	0.020		0.032	0.014	0.041	0.041	0.029			2.936	0.037
STD DEV	0.180	0.146	0.142		0.178	0.118	0.203	0.203	0.167				



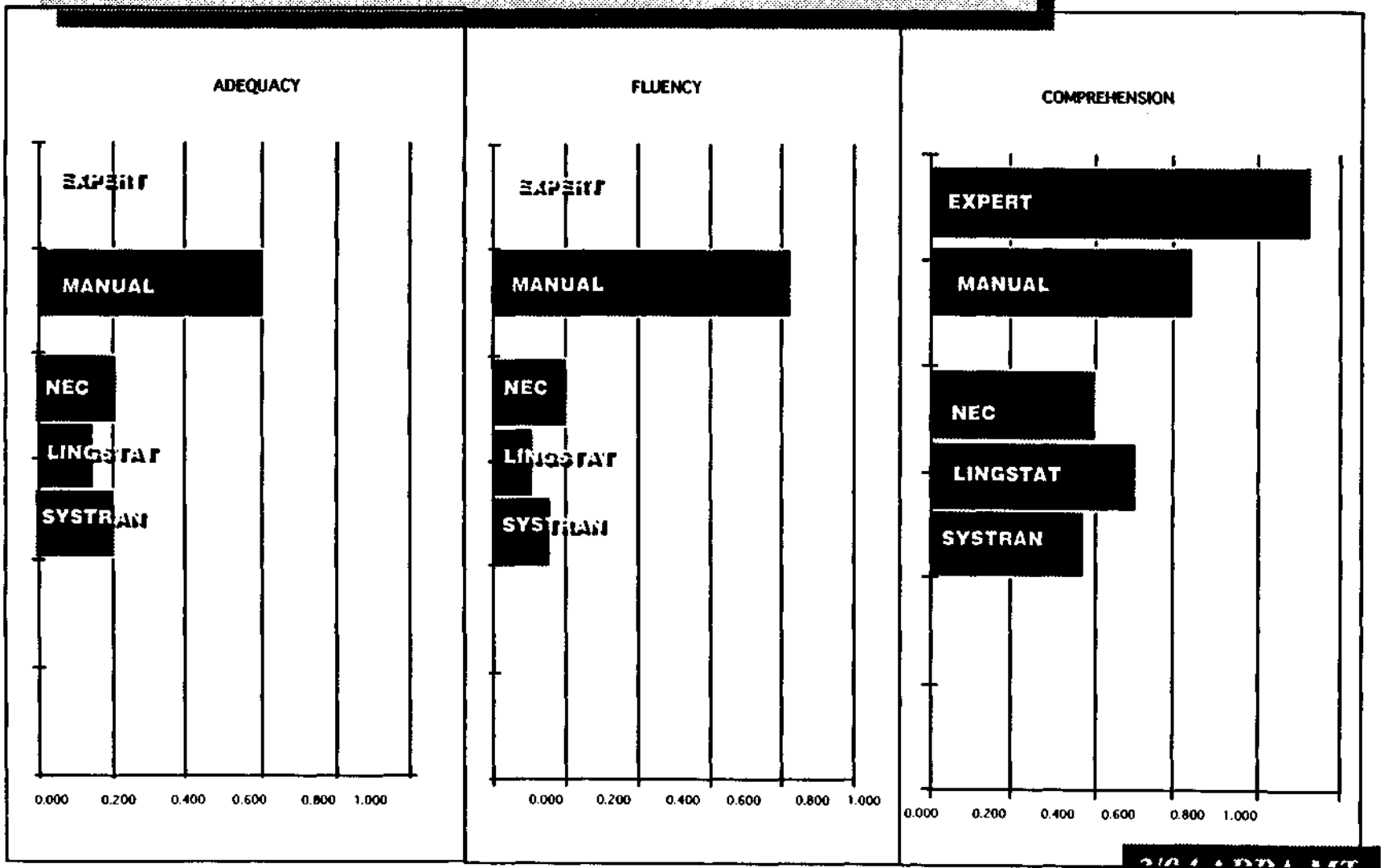
French FAMT Systems Results



Spanish FAMT Systems Results

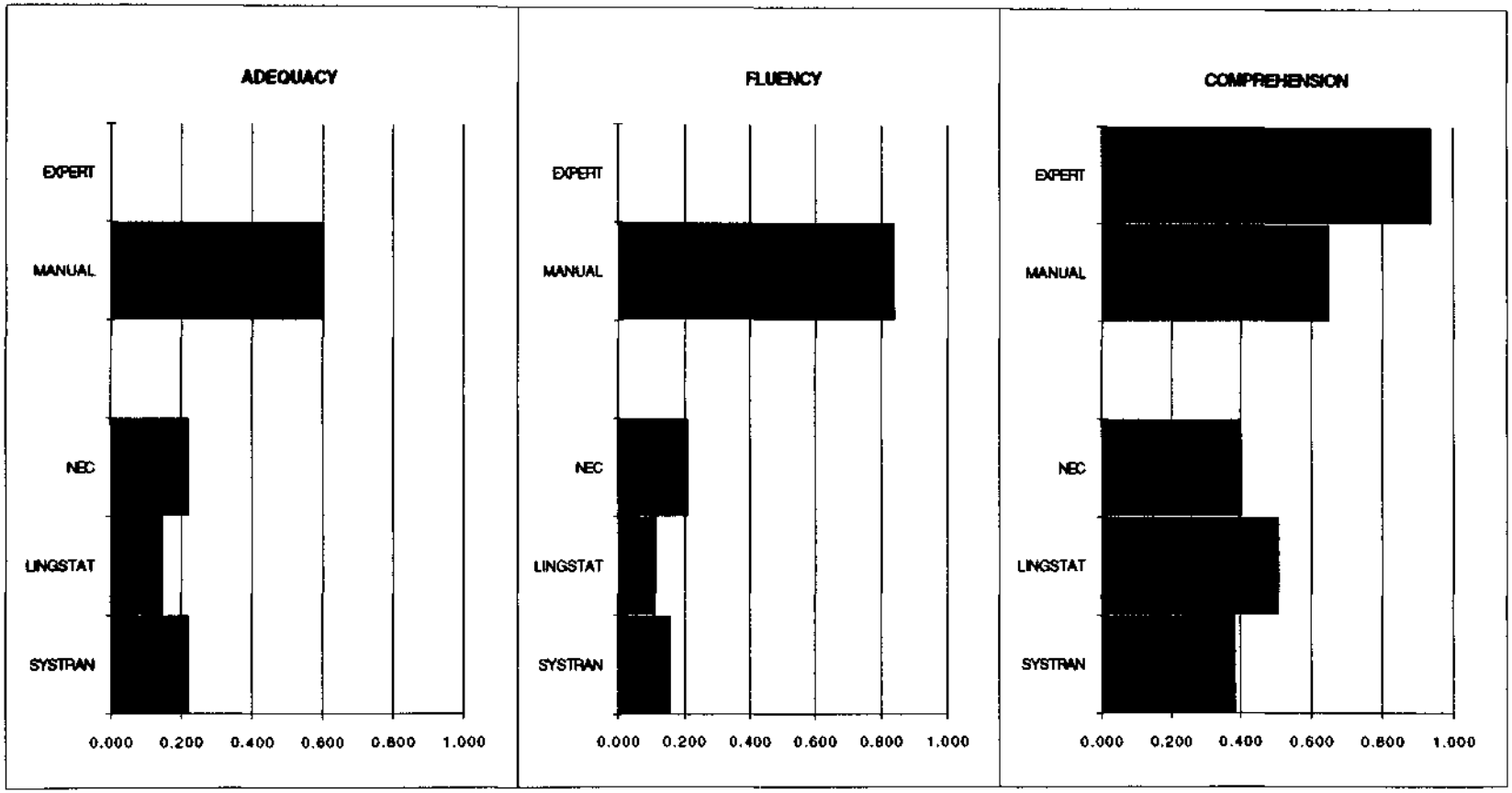


Japanese FAMT Systems Results



FAMT EVALUATION FOR JAPANESE, 1Q94

	LINGSTAT	NEC	MANUAL	EXPERT	SYSTRAN	MEAN	STD DEV	VARIANCE	Total F Ratio	AVG STD DEV
COMPREHENSION	0.509	0.404	0.649	0.939	0.386	0.577	0.228	0.052	3.669	0.054
VARIANCE	0.093	0.068	0.062	0.015	0.078	0.063				
STD DEV	0.305	0.260	0.248	0.124	0.279	0.243				
FLUENCY	0.114	0.211	0.841		0.159	0.331	0.342	0.117	F-RATIO	AVG STD DEV
VARIANCE	0.013	0.028	0.018		0.011	0.018			29.853	0.029
STD DEV	0.113	0.166	0.135		0.107	0.130				
ADEQUACY	0.147	0.223	0.604		0.223	0.299	0.206	0.043	F-RATIO	AVG STD DEV
VARIANCE	0.015	0.033	0.038		0.016	0.025			7.511	0.035
STD DEV	0.122	0.182	0.195		0.125	0.156				



Summary

- All three evaluations have value
  - Did calibrate underlying MT technologies
  - But not intended as diagnostic of system value/effectiveness
- HAMT - - NOT
- Valuable portability exercise
  - Gracious participation by production systems
- A healthy technology evaluation paradigm for both ARPA and the MT industry