MUC-3 LINGUISTIC PHENOMENA TEST EXPERIMENT

Nancy Chinchor, Ph.D.
Science Applications International Corporation
10260 Campus Point Drive, M/S 12
San Diego, CA 92121
(619) 458-2728

INTRODUCTION

The evaluation of data extraction systems can be supplemented by determining performance of the systems on a representative selection of linguistic phenomena. The experiment performed as part of the MUC-3 evaluation was aimed at determining whether linguistic phenomena can be tested in isolation for state-of-the-art data extraction. Although not all of the methods of data extraction used by the participants in MUC-3 directly process linguistic phenomena, the methods are dealing with those phenomena in some manner because they are present in the input. Phenomena testing in MUC-3 is testing according to the characteristics of the messages rather than characteristics of the systems processing those messages. In order to determine the validity of phenomena testing at the current level of system performance, an experiment was run. The results of the experiment indicate that linguistic phenomena are isolatable and that performance for linguistic phenomena can be measured using the MUC-3 scoring system.

DESIGN

The problem is to devise a test to measure the performance of a data extraction system with respect to a single linguistic phenomenon. The experiment took several approaches to devising such a test to determine whether the phenomenon had been isolated. The design of the experiment required the choice of a linguistic phenomenon frequently appearing in the messages and critical to the template fill task.

The slots from phrases exhibiting the phenomenon would be scored and compared to the overall scores. If there was no correlation with the overall scores, then the possibility that overall scores were fully determining the phenomenon's scores would be eliminated and isolating the phenomenon would be possible. The slots filled from the sentences were scored and compared to the overall scores as well as the scores obtained for the slots filled from the phrases. If the scores for the sentences correlated more closely with the scores for the phrases than they did with the overall scores, then it would be more likely that the scoring was isolating the phenomenon. Processing of the phrase can have an effect on the processing of the entire sentence. If the results for the phrases and the sentences coincide, then it would be feasible to use scores for slots from entire sentences for future phenomena testing.

Slots from phrases exhibiting well-defined subsets of the phenomenon would be scored and compared to each other. The results of the comparisons that can be predicted or explained would give us more confidence that we have isolated the phenomenon.

Altered messages would be produced without the phenomenon for purposes of a "minimal pair" type test. The responses would be scored for slots filled from the phrases that formerly constituted the phenomenon. The scores would be compared to the scores for responses to the original messages containing the phenomenon. The comparison would provide more information concerning the success of isolating the phenomenon.

METHOD

Choice of Phenomenon

Apposition was chosen as the linguistic phenomenon because of its frequency of occurrence in messages and its criticalness for slot fills. An example of an appositive from the test corpus is "David Lecky, Director of the Columbus School." There were approximately 60 sentences in the test corpus containing one or more appositives which were critical to slot fills. Preliminary phenomena testing for three other phenomena occurring with varying frequencies suggested that a frequency of 20 was adequate for testing purposes. With more than 60 instances of apposition, subdividing the set for testing well-defined subsets would still leave adequate numbers in the subdivisions. Also, there many more cases of appositives which affected slot fills, but could not be included in the testing because there were other sources for the slot fills elsewhere in the message. This high frequency of occurrence of apposition in the messages suggests that it is a phenomenon which systems must handle in some way.

Definition of Apposition

The examples used from the test messages are all cases of noun phrases in apposition. Among linguists, there is variation in the liberality with which the term *apposition* is used. According to Quirk et al [1], apposition meeting the following three conditions is *full apposition*:

a. each of the appositives can be separately omitted without affecting the acceptability of the sentence;

b. each fulfills the same syntactic function in the resultant sentence; and

c. there is no difference between the original sentence and either of the resultant sentences in extralinguistic reference.

An example of full apposition is the following from test message TST2-MUC3-0004.

> JOSE PARADA GRANDY, THE BOLIVIAN POLICE CHIEF, TOLD EFE THAT AN UNIDENTIFIED PERSON STEPPED OUT OF A VEHICLE AND PLACED A PACKAGE IN ONE OF THE PLANT POTS ON JUAN DE LA RIVA STREET, A FEW METERS FROM THE U.S. EMBASSY IN DOWNTOWN LA PAZ.

"Jose Parada Grandy" and "the Bolivian Police Chief" are in full apposition because they each can be omitted resulting in the following acceptable sentences, they each are the subject in those sentences, and all three sentences have the same extralinguistic reference.

JOSE PARADA GRANDY TOLD EFE THAT AN UNIDENTIFIED PERSON STEPPED OUT OF A VEHICLE AND PLACED A PACKAGE IN ONE OF THE PLANT POTS ON JUAN DE LA RIVA STREET, A FEW METERS FROM THE U.S. EMBASSY IN DOWNTOWN LA PAZ.

THE BOLIVIAN POLICE CHIEF TOLD EFE THAT AN UNIDENTIFIED PERSON STEPPED OUT OF A VEHICLE AND PLACED A PACKAGE IN ONE OF THE PLANT POTS ON JUAN DE LA RIVA STREET, A FEW METERS FROM THE U.S. EMBASSY IN DOWNTOWN LA PAZ.

Partial apposition occurs when the three conditions are not all met. An example of partial apposition not meeting condition (a) appears in test message TST2-MUC3-0100.

THE BRAZILIAN EMBASSY IN COLOMBIA HAS CONFIRMED THE RELEASE OF REDE GLOBO JOURNALIST CARLOS MARCELO WHO WAS KIDNAPPED BY COLOMBIAN ARMY OF NATIONAL LIBERATION GUERRILLAS.

The difference between full and partial apposition in this case is trivial requiring only the addition of a determiner to "Rede Globo journalist" to make the sentence omitting "Carlos Marcelo" acceptable. Partial appositives that were omitted from the phenomenon testing were cases of appositives containing "also" and "alias." These were omitted because of their adverbial nature.

Another gray area in choosing examples concerns titles. Quirk et al makes the distinction between apposition and institutionalized titles. The authors show the range from apposition in "critic Paul Jones" to full title in "Mr. Porter" with the following examples:

a. critic Paul Jones
   the critic Paul Jones (with appositives, a preposed determiner is normal but not with titles)
   Paul Jones the critic (with appositives, postposition with "the" is more normal than preposition without "the" whereas the opposite is true for titles that allow postposition)
   the critic (appositives and most titles can be used without the proper nouns and with determiners)
   ?critic (vocative) (most titles and some appositives can be used as vocatives)

b. Farmer Brown
   the farmer Brown
   ?Brown the farmer
   the farmer
   farmer (vocative)

c. Brother George (family)
   my brother George/ ?the brother George
   *George the brother
   the brother
   brother (vocative)

d. Professor Brown
*the professor Brown
?Brown the professor
the professor
professor (vocative)

e. Dr. Smith (Ph.D.)
*the doctor Smith
*Smith the doctor
*the doctor
doctor (vocative)

f. Mr. Porter

*the Mr. Porter	(with titles, a preposed determiner is not normal)
*Porter the mister	(postposition with "the" is not allowed here)
*the mister	(some titles cannot be used without the proper nouns and with determiners)
*mister (vocative) (substandard)	(most titles can be used as vocatives)

In the MUC-3 messages, the appositives and titles are distinguished by the tests above with the cut-off between (3) and (4). For example, "Colonel," "Senator," and "Ambassador" are titles because the following judgments are similar to those for "Professor" above:

Colonel Heriberto Hernandez

*the Colonel Heriberto Hernandez	(with titles, a preposed determiner is not normal unless the noun phrases are modified restrictively)
?Heriberto Hernandez the Colonel	(with titles that allow postposition, preposition without "the" is more normal than postposition with "the")
the Colonel	(appositives and most titles can be used without the proper nouns and with determiners)
Colonel (vocative)	(most titles can be used as vocatives)

However, "student" and "peasant" are considered appositives because of the following pattern similar to the pattern for "critic" above:

student Mario Flores
the student Mario Flores
Mario Flores the student
the student
?student (vocative)

Judgments may vary. One possible questionable inclusion as an appositive is the phrase "Attorney General." My judgments follow:

Attorney General Roberto Garcia Alvarado
the Attorney General Roberto Garcia Alvarado
Roberto Garcia Alvarado the Attorney General
the Attorney General
Attorney General (vocative)

An attempt was made to limit the appositives used in the testing to those most likely to be agreed upon as appositives while still maintaining a reasonable number of examples.

Construction of the Test Sets

The message sentences containing appositives were extracted from the messages for analysis. The examples were put in a file for distribution to the participants to assist in analysis of their results. This file contained information concerning the categorization of the appositives and the slots affected by the appositioned noun phrases and the entire sentence.

The appositives were categorized as postposed versus preposed and simple versus complex. An example of a postposed appositive is "Jose Parada Grandy, the Bolivian Police Chief" and an example of a preposed appositive is "Rede Globo journalist Carlos Marcelo." The subdivision of the appositives according to their complexity was done subjectively based on internal structure and the context. Both "Jose Parada Grandy, the Bolivian Police Chief" and "Rede Globo journalist Carlos Marcelo" were considered simple. Any complexity in an example, such as conjunction within the appositive, a missing comma, or a comma inside double quotes, put that example in the complex category. Probably the most complex appositioned noun phrase in the corpus was in apposition to "peasants" in TST2-MUC3-0036. The misspelling "Colonal" is part of the message.

> THE PEASANT COMMUNAL ASSOCIATION, ACC, CONTINUES TO DEMAND THE RELEASE OF PEASANTS BARTOLO RODRIGUEZ, WHO WAS CAPTURED ON 27 JANUARY, AND [NAME INDISTINCT] CAPTURED ON 2 FEBRUARY BY TROOPS OF COLONAL ORLANDO MONTANO OF THE 6TH INFANTRY BRIGADE.

The most important and difficult activity in constructing phenomena tests is to determine the individual slots that could only be filled from the phrase containing the phenomenon being tested. The slots that could only be filled by the information in the appositioned noun phrases as well as in the sentences containing those appositioned noun phrases were noted. The configuration option files for the scoring system were constructed to score just those slots directly affected by the presence of an appositive. Slots that could have been filled from any other phrase/sentence not containing an appositive were excluded from the scoring. This step in the test construction is the most likely point where human error can intrude.

For the purposes of running the "minimal pair" test, a modified version of the message file was produced. The messages were altered to contain simple sentences expressing the equivalence of the appositioned noun phrases in cases where the appositioned noun phrases directly affected at least one slot in the template fill. The appositive no longer appeared in the original sentence. For example,

> THE BRAZILIAN EMBASSY IN COLOMBIA HAS CONFIRMED THE RELEASE OF REDE GLOBO JOURNALIST CARLOS MARCELO WHO WAS KIDNAPPED BY COLOMBIAN ARMY OF NATIONAL LIBERATION GUERRILLAS.

was replaced by

> THE BRAZILIAN EMBASSY IN COLOMBIA HAS CONFIRMED THE RELEASE OF CARLOS MARCELO WHO WAS KIDNAPPED BY COLOMBIAN ARMY OF NATIONAL LIBERATION GUERRILLAS. CARLOS MARCELO IS A REDE GLOBO JOURNALIST.

The "minimal pair" test was voluntary because it required a separate run of the data extraction systems on the modified messages.

The scoring of the appositive tests was diluted somewhat by the allowance in the scoring guidelines for partial credit to be given when the key contains a complete proper name and the response contains only the identifying part of the name. It was typical of the appositioned noun phrases that they were the place where the full name of the person was introduced with only part of the name being used from then on for reference. A previously undetected bug in the scoring system caused one template not affected by apposition to be scored instead of another template that was affected by apposition. For phrases, only 2 slots out of a possible 66 slots (3%) were potentially affected; for sentences, 9 slots out of a possible 198 slots (4.5%) were potentially affected.

HYPOTHESES

The intent of the testing was to discover whether the scoring isolated the phenomenon of apposition. Each of the following hypotheses was proposed and tested in order to uncover evidence of isolation of the phenomenon.

> Hypothesis 1. The systems should score differentally on the appositives (both phrasally and sententially) than they did on the overall testing.
>
> Hypothesis 2. The systems should score higher on the simpler appositives.
>
> Hypothesis 3. The systems should score differently on the postposed and preposed appositives. It was not possible to hypothesize which score would be higher. Although postposed appositives are more prototypical and have indications they are appositives such as commas or dashes, preposed appositives lend themselves to treatment as adjectives.
>
> Hypothesis 4. The systems should score higher on their responses to the messages where simple sentences were substituted for appositives.

RESULTS

The recall and precision scores for the appositive tests appear in Table 1. Table 2 contains the scores based on the single measure calculated by multiplying recall times precision.

Analysis of Results

Hypothesis 1 asserts that the apposition results are independent of the overall performance of the systems. To determine the validity of Hypothesis 1, scatter plots were made of overall recall versus precision scores for a test run under comparable conditions (Figure 1), the appositive scores for phrases (Figure 2), and the appositive scores for sentences (Figure 3). Comparing Figures 1 and 2 shows that the scores for apposition are significantly different from the overall scores. The performance by systems on apposition is largely independent of their overall scores. The same conclusion can be drawn for the appositive scores for sentences by comparing Figures 1 and 3.

Appositive Results

Site	App R	App P	Sen R	Sen P	Easy R	Easy P	Hard R	Hard P	Post R	Post P	Pre R	Pre P
ADS	0	0	1	40	0	0	0	0	0	0	0	0
BBN	20	31	26	42	35	41	12	21	18	27	23	35
GE	32	48	22	54	48	50	20	47	27	36	32	73
GTE	2	30	3	32	0	0	2	33	3	33	0	0
HU	20	16	19	29	30	14	14	25	29	23	8	9
ITP	11	54	7	66	7	66	10	43	12	67	10	43
LSI	2	14	3	28	3	12	1	25	3	33	0	0
MDC	20	29	14	28	22	38	18	24	27	26	13	40
NYU	32	62	21	57	42	77	25	52	23	50	42	74
PRC	8	42	10	48	20	57	1	25	12	67	3	25
SRI	25	63	17	59	35	58	19	67	26	57	23	70
SYN	0	0	0	0	0	0	0	0	0	0	0	0
UM	43	77	32	65	68	84	32	71	45	68	40	92
UN	2	25	6	40	2	25	2	25	4	25	0	0
UNI	15	40	11	46	32	54	6	21	14	45	18	37

Table 1: The sites reported recall and precision scores for the appositive phrases, the sentences containing appositives, the easy and hard appositives, and the post-posed and preposed appositives.

Single Appositive Measures

Site	Easy R X P	Hard R X P	Post R X P	Pre R X P
ADS	0	0	0	0
BBN	1435	252	486	805
GE	2500	940	1008	2336
GTE	0	66	99	0
HU	420	350	667	72
ITP	462	430	804	430
LSI	36	25	99	0
MDC	836	432	702	520
NYU	3234	1300	1150	3108
PRC	1140	25	804	75
SRI	2030	1273	1482	1610
SYN	0	0	0	0
UM	5712	2272	3060	3680
UN	50	50	100	0
UNI	1728	126	630	666

Table 2: The single measure scores were calculated for comparing easy versus hard and postposed versus preposed appositives.

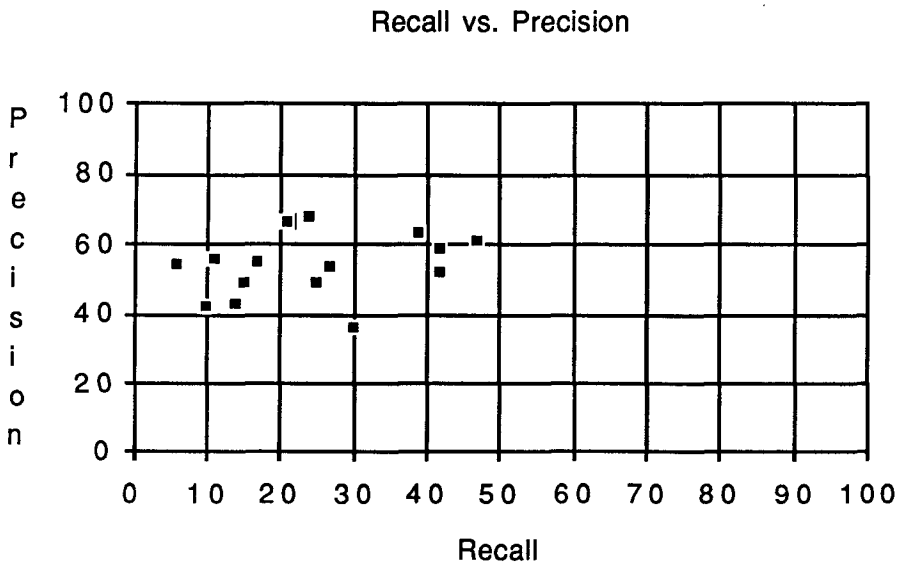

Figure 1: A scatter plot shows the scores for overall recall versus precision in a test run under comparable conditions.

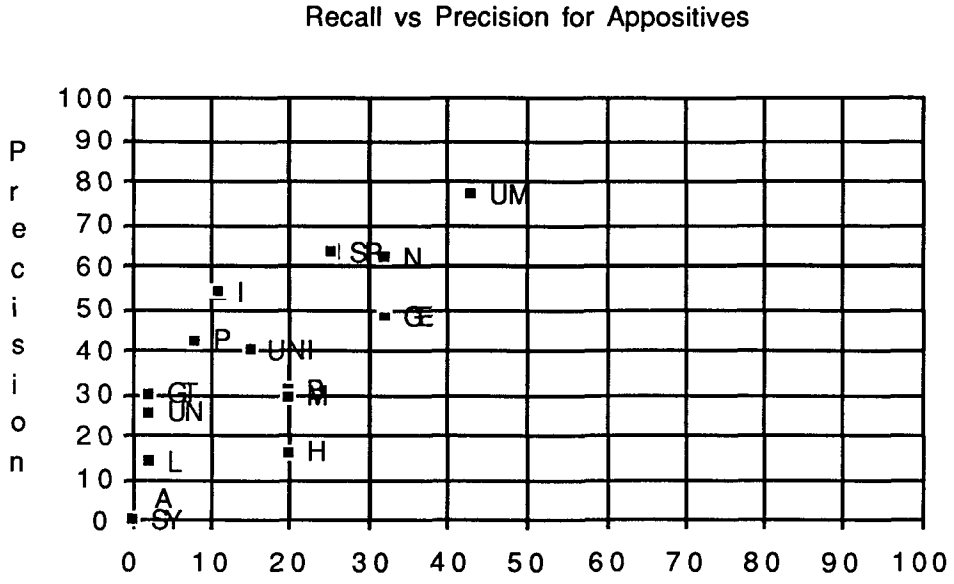

Figure 2: The recall versus precision scores for appositive phrases shows that the performance is different from the overall performance.

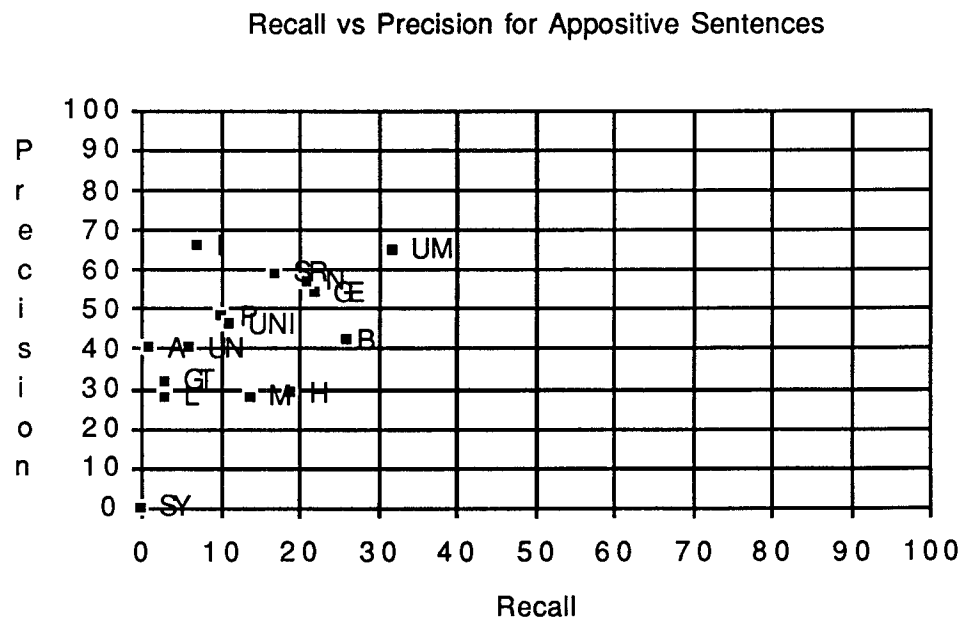

Figure 3: The recall versus precision scores for appositive sentences are more like the scores for phrases than like the overall scores.

The scatter plots for appositives scored from phrases and sentences in Figures 2 and 3, respectively, are more comparable to each other than to the overall scores suggesting that the use of information from sentences could be a valid test of performance on a phenomenon. Further analysis illustrated in Figures 4 and 5 shows that the scores for appositives and sentences containing appositives parallel each other for both recall and precision. These parallelisms affirm that material from sentences containing a phenomenon can be used for testing that phenomenon and also indicate that we may be isolating the phenomenon.

Hypothesis 2 asserts that the systems will score higher on the simpler appositives than on the more complex ones. The scores for recall are remarkably higher for the easy appositives as opposed to the harder appositives as shown in Figure 6. Figure 7 shows a less clear trend for the precision scores. The single measure of recall times precision, however, shows an unmistakable trend of systems scoring more highly for the easier appositives. These results give us confidence that we are isolating the phenomenon of apposition.

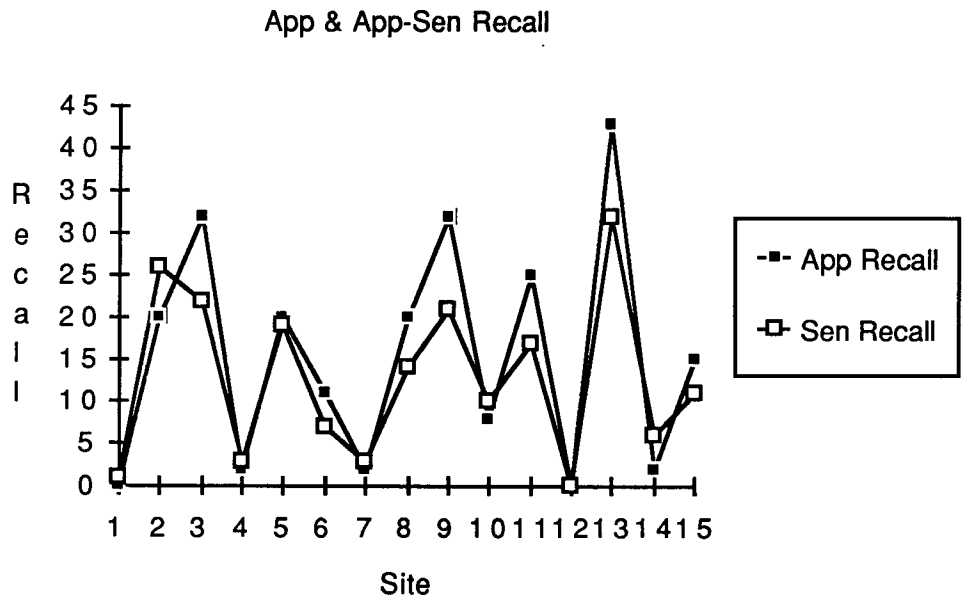

Figure 4: The recall scores for appositives and sentences containing appositives correlate with each other.

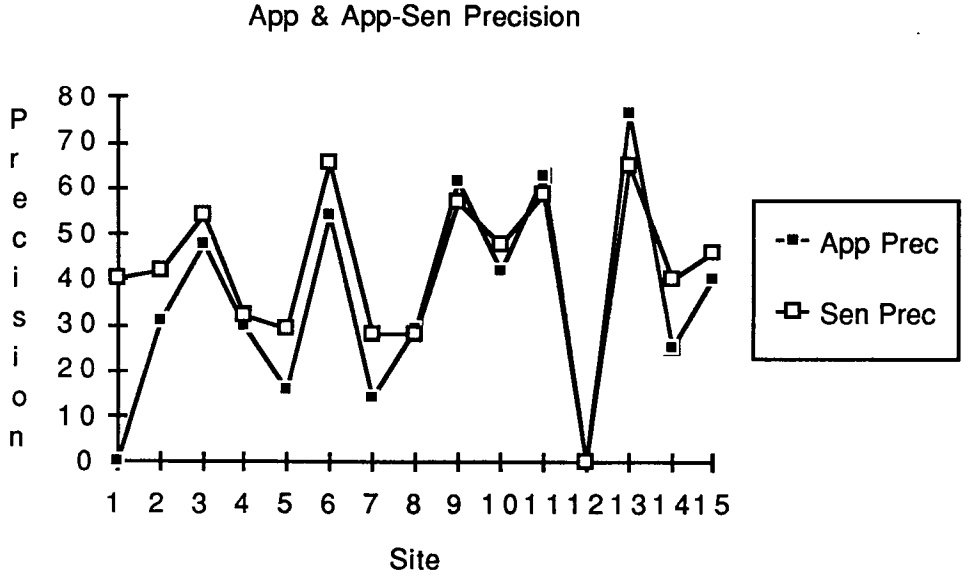

Figure 5: The precision scores for appositives and sentences containing appositives correlate with each other.

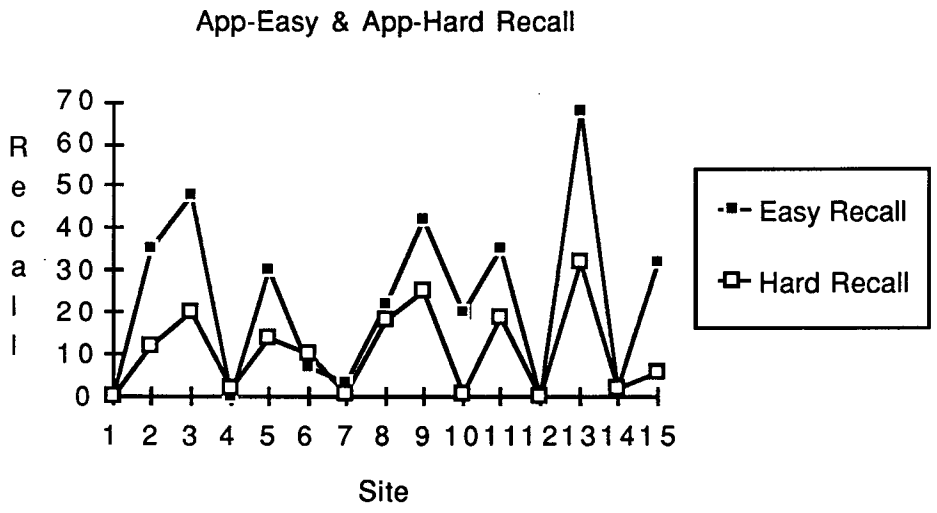

Figure 6: The recall scores for the easy appositive phrases are generally higher than those for the harder phrases.

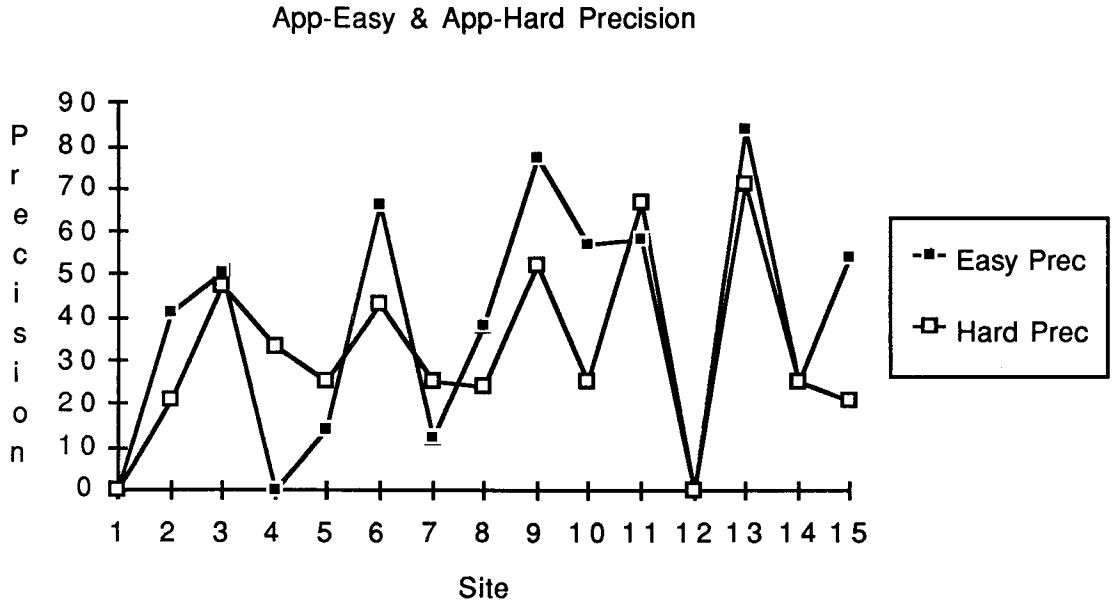

Figure 7: The precision scores show a tendency to be higher for the easier appositions.

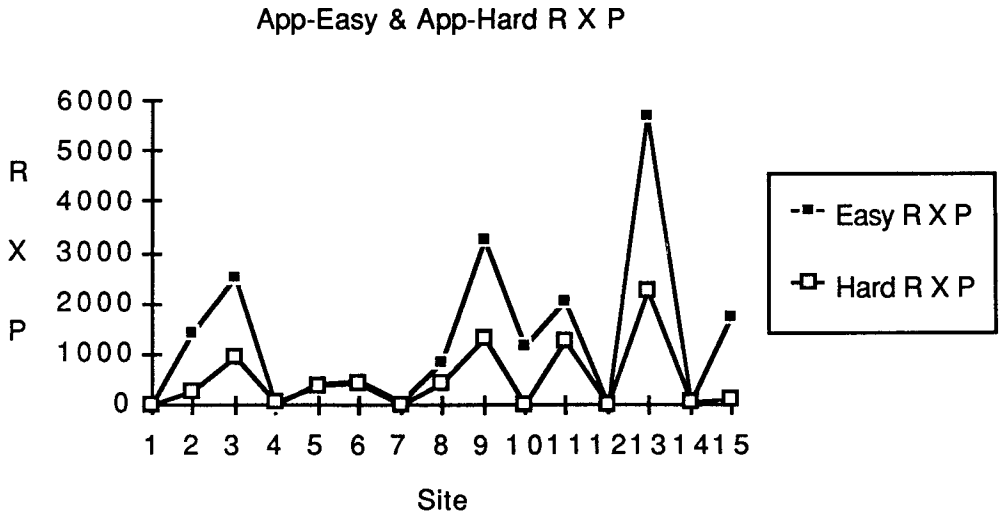

Figure 8: The single measure scores of recall times precision are higher for the easier appositions than for the harder ones.

The inability to predict whether postposed or preposed appositives would score higher was actually supported by the data. Hypothesis 3 was born out in that the systems did score differently on the two types of appositives. There was no clear trend in the results as to which kind of apposition was easier. The recall, precision, and single measure scores are shown in Figures 9 through 11. Notice that the results were predicted providing further evidence that the phenomenon of apposition is being isolated. It would be interesting to look at the methods of processing the two types of appositives for each of the systems to see why their scores are as they are.

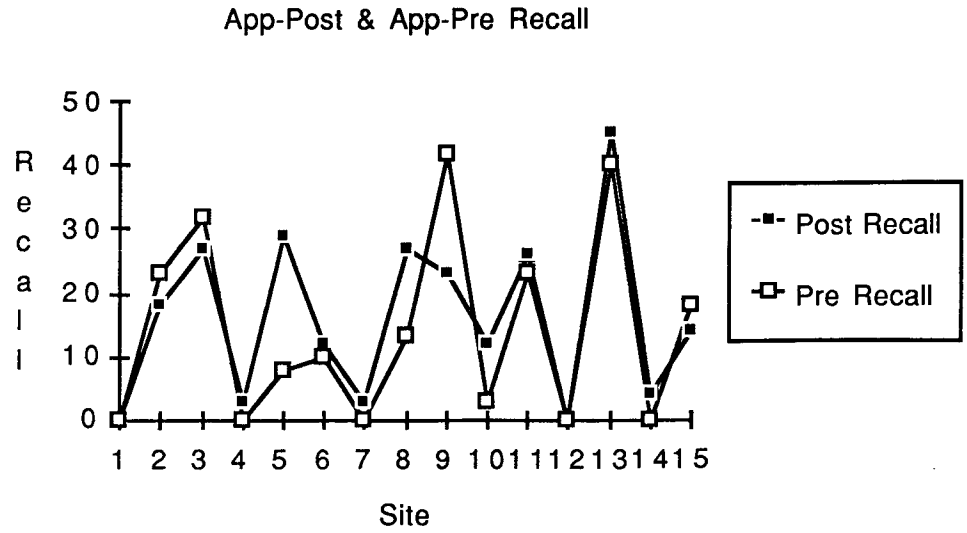

Figure 9: The recall scores for postposed and preposed appositives show differences in the scores but no clear trend as to which is easier to process.

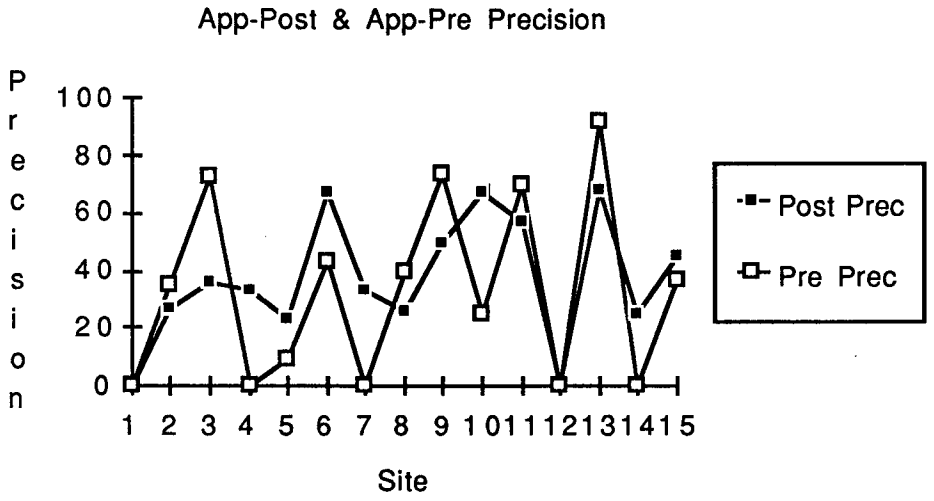

Figure 10: The precision scores for postposed and preposed appositives show differences in the scores but no clear indication as to which is easier to process.

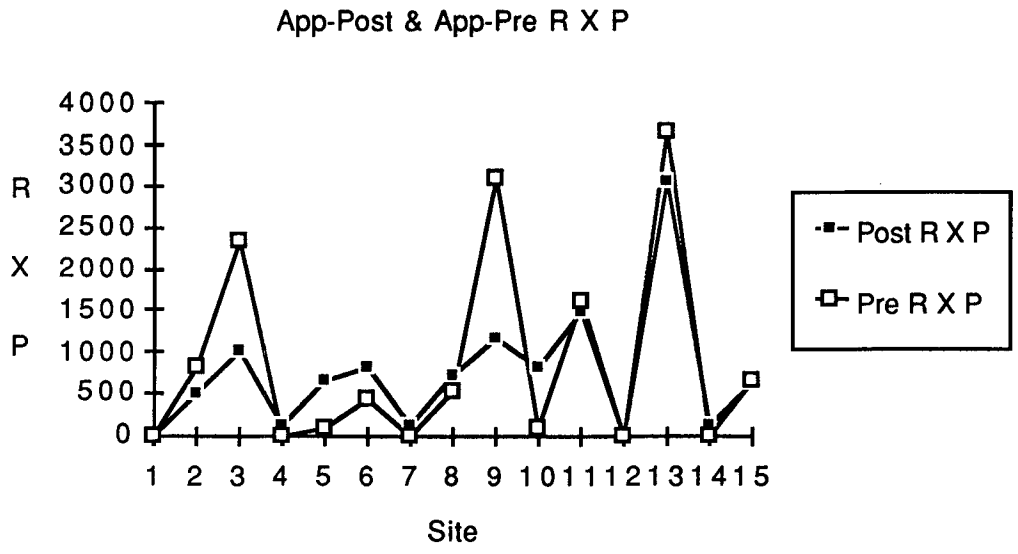

Figure 11: The single measure of recall times precision for postposed and preposed appositives shows that the systems score differently on the two but neither is consistently easier.

Hypothesis 4 predicts that the systems will score higher for the messages containing simple sentences in place of the appositives. Two sites volunteered to run this part of the test and they both contradicted the hypothesis. Their scores are shown in Table 3 alongside their scores for the messages containing the appositioned phrases. On further analysis, it was found that the introduction of the simple sentences made the task more complex in both cases. Apparently, the appositioned noun phrases convey the information more simply than a separate sentence containing a copula and requiring reference resolution. The systems, for various reasons, tended

not to use the information in the separate sentence. The recall scores are thus lower. The precision scores are somewhat affected. The results show an explanable effect on the scores lending further credence to the claim that the apposition phenomena is being isolated.

VOLUNTARY

Site	Recall	Precision	App R	App P
NYU	28	53	32	62
UMASS	38	68	43	77

Table 3: The recall and precision scores for the voluntary "minimal pair" test for the messages without apposition and the messages with apposition show an effect of modifying the appositioned noun phrases.

CONCLUSIONS

In summary, the systems scored differently on the appositives than they did on the overall testing suggesting that the testing may be isolating the phenomenon of apposition. The systems scored similarly on the slots filled from phrases containing appositives and sentences containing appositives suggesting that information from sentences could be used to test phenomena. Because the processing of apposition can affect the processing of the entire sentence, the parallel results in these scores further suggests that the phenomenon of apposition is being isolated. The systems scored markedly higher on the simpler appositives as opposed to the more complex ones. These results are perhaps the strongest evidence that it is possible to isolate the phenomenon of apposition by scoring slot fills. The systems scored differently on the postposed and preposed appositives. It would be interesting to look at the methods employed by each system with respect to these classes of appositives. It was predicted that neither class would be clearly easier. The fact that this prediction was correct provides strong support for the claim that apposition is being isolated. The systems scored lower on their responses to the messages where simple sentences were substituted for appositives. The effect on the scores, although unexpected, still supports the isolatability of apposition. In some of the more well-defined trends, the anomalies noticed are often for the lower scoring systems. However, the systems are scoring highly enough overall at this stage of development for the phenomena scores to be meaningful. In conclusion, there are strong indications that the phenomenon of apposition has been isolated by the testing and that performance on apposition can be scored using the MUC-3 scoring system.

Further Research

Further work in phenomena testing should now be focused on carefully developing a representative selection of phenomena tests for the messages. The evaluation of data extraction systems can be enhanced by determining performance of the systems on these linguistic phenomena. Phenomena testing should be done at various linguistic levels including the word level, phrase level, sentence level, intersententially, and the level of discourse. Testing according to the linguistic characteristics of the messages would encourage the data extraction systems to improve capabilities applicable to other domains.

REFERENCES

[1] Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J., *A Grammar of Contemporary English* (London: Longman Group Limited, 1984).

PART II: TEST RESULTS AND ANALYSIS (SITE REPORTS)

The papers in this section were prepared by each of the fifteen sites that completed the MUC-3 evaluation. The papers are intended to provide the reader with some context for interpreting the test results, which are presented more fully in appendices F and G of the proceedings. The sites were asked to comment on the following aspects of their MUC-3 experience:

* Explanation of test settings (precision/recall/overgeneration) and how these settings were chosen

* Where bulk of effort was spent, and how much time was spent overall on MUC-3

* What the limiting factor was (time, people, CPU cycles, knowledge, ...)

* How the training of the system was done
  - What proportion of the training data was used (and how)
  - Whether/Why/How the system improved over time, and how much of the training was automated

* What was successful and what wasn't, and what system module you would most like to rewrite

* What portion of the system is reusable on a different application

* What was learned about the system, about a MUC-like task, about evaluation