

PRACTICAL ASPECTS OF CORPUS TAGGING

Steve Seegmiller and Eileen Fitzpatrick
Montclair State University

1. Introduction

At Montclair State University, we are in the process of creating an on-line, publicly accessible database of learner English. The database will be called MELD, for the Montclair Electronic Language Database, and we expect it to be an ongoing project, with additional data being added continuously.

The goals of the project are to create a database that will be used by students and faculty in Montclair State University's own applied linguistics program, but in addition we hope that the database will be useful to researchers on second-language acquisition, for computational linguists building grammar checkers, student editing aids, and other tools; and to teachers seeking materials for use in the classroom.

At the present time, we have a number of essays by students learning English at Montclair State University that have been converted to ASCII text files and placed on our server. Eventually, the texts will be available in both the plain, ASCII, untagged form, and separately in a form in which the errors have been tagged or annotated. We will return to a discussion of our system of tagging later in the paper.

Our long-range goals include adding data from other languages, other institutions, and other countries, as well as having a web-based interface that will allow the user to carry out various kinds of searches and downloads.

A corpus of language-learner data is a valuable tool for researchers. For example, a comparison of the differences between corpus of second-language-learner (L2) writing and that of native-language (L1) writing can be an invaluable source of information for people preparing teaching materials or studying the process of L2 acquisition. The work in Granger (1998), for example, shows differences in phrase choice (Milton), differences in complement choice (Biber and Reppen), and differences in word choice, sentence length, and lexical variation (Meunier).

In general, corpora that are annotated for errors are more useful, and provide more data, to researchers. Various approaches to error tagging have been proposed; Polio (1997), for example, reviews measures of writing accuracy in L2 research. These measures include holistic scales, counts of error free units, counts of errors, and identification and classification of errors. Only a few of the studies cited by Polio report interrater reliability or consistency between annotators. Among the research that reports reliability of errors identified, Kepner (1991) reports a .97 percent reliability, Zhang (1987) .85 percent, and Bardovi-Harlig and Bofman (1989), which identifies and

classifies errors, .88 percent. Polio's own study of interrater reliability, which is the most explicit in its description of the methods used, reports .74 percent agreement in classifying error type.

2. Tagging Methods.

In this paper, we describe the system of tagging that we use in the MELD database, and some practical considerations related to the process of tagging the data in the database. There are two quite different approaches to error tagging: **classification**, in which taggers categorize writers' errors according to a predetermined list of error types, and **reconstruction**, in which taggers identify the errors in the texts and replace them with the correct form.

Each of the two approaches to error tagging has its advantages and disadvantages. Researchers using the classification approach, for example, report very high levels of agreement between taggers [refs, data]. In addition, since classification produces data that are easily counted, statistical analysis is easily accomplished. However, there are three main disadvantages to classification. First, every error must be assigned to a single category, even if there is more than one possible way of interpreting the error. For example, an error like

(1)... computers and television are great aid in searching and helping you study...

(which was taken from one of the essays in our database) might be interpreted in either of two ways: as missing an article (... a great aid ...) or as an error in the number of a noun (... great aids ...). We would not want to say that one of these ways of classifying the error is correct and the other wrong; rather, we would say that these are two alternative ways of correcting the error. A classification system of tagging forces the tagger to decide between the two interpretations, and thus tends to lead to arbitrary and inconsistent classifications.

A second disadvantage to the classification approach to tagging is that error types that were not anticipated when the categories were devised will be difficult to classify. They will either be arbitrarily classified or placed in a (perhaps large) "miscellaneous" category.

The third disadvantage to classification is that any classification system is at least partly theory-dependent, and categories established on the basis of one researcher's theory may be irrelevant to another.

Reconstruction avoids all of these problems. Since reconstruction does not depend on a predetermined list of error types, it provides unlimited flexibility in classifying error types and it allows the possibility that a single error can be reconstructed on more than one way. Reconstruction has the added advantage that there is no classification system for the taggers to learn so less training is necessary. It also provides the kind of data that

is more useful for text-processing applications. For example, if one wanted to write an editor to help students revise their own writing, it would be much more useful to have a body of errors and proposed corrections than to have a list of how many times a given error type occurred. The primary disadvantage to reconstruction is that, since it does not provide numerical or countable data directly, it is harder to analyze statistically.

3. Tagging Experiments.

After weighing the advantages and disadvantages of each type of tagging, we have decided to adopt the reconstruction approach. Our next task is to adopt a procedure for tagging that our taggers (our own graduate and undergraduate students) can use. The obvious question that arises is whether one tagger is sufficient for each essay, or whether multiple taggers must tag each item. In previous work (Fitzpatrick and Seegmiller 2000) we reported on two experiments involving three taggers. One tagger participated in both experiments. In Experiment A, the data consisted of student essays written in a fixed time period of 20 minutes on an assigned topic. The two taggers worked independently but discussed their tagging after each essay. In Experiment B, two taggers tagged one training set, discussed their results, tagged a first experimental set, discussed the second set of results, and then tagged a final set of essays. The results of these two experiments are shown in Table 1. For Experiment B, Test 1 and Test 2 are show the results before and after the second discussion of results.

Table 1.
Results of Experiments A and B

	<i>Precision</i>	<i>Recall</i>
<i>Experiment A</i>	0.85	0.81
<i>Experiment B</i>		
<i>Test 1</i>	0.84	0.73
<i>Test 2</i>	0.9	0.76

Precision and **recall** are standard measures borrowed from computational linguistics, specifically from informational retrieval, where they are used to measure the performance of a computer against that of a human. We will describe these measures later in more detail, along with a third measure we used, **reliability**. Suffice it to say now that scores of .80 and above are considered to indicate adequate measures of agreement between taggers. As we see in Table 1, the procedures of Experiments A and B are generally encouraging. showing both a high degree of agreement between taggers, and an increase in agreement between Tests 1 and 2 for Experiment B.

The question we faced next was whether we could achieve similarly positive

results with a group new group of taggers made up of our own graduate students. As the database develops, we expect to use our own students to enter and tag the data. If we can devise a training method that results in agreement scores of .80 or above for these taggers, then we can have one student tag each essay and be confident that the results achieved will not be substantially different than if we had used more than one tagger. On the other hand, if our student taggers cannot reach a high level of agreement in their tagging, we will have to adopt some procedure for dealing with the disagreements.

For our third experiment, we recruited three experienced teachers of ESOL. We provided them with a brief initial orientation and then asked them to tag the first set of ten essays tagged, totaling approximately 5,000 words. We then reviewed their tags, noted the discrepancies, and held two lengthy meetings of about two hours each to discuss discrepancies in the tags and to establish a set of guidelines for them to follow when tagging the second set of essays. These guidelines included the following:

1. Make a change only if the original is incorrect, not if it correct but could be stated in a better way.
2. Do not insert a comma after an introductory prepositional phrase, but do insert one after an introductory adverbial clause.
3. Lexical substitutions may be made to eliminate inappropriately colloquial lexical items (e.g. "child" for "kid") but not otherwise (e.g. not "select" for "choose").

After these meetings, the taggers annotated a second set of 12 essays, also totaling approximately 5,000 words. Table 2 gives the raw agreement data for each of the three pairs of taggers. The three taggers are designated by the initials A, B, and C, and Table 2 shows the pairwise comparisons among them. "Same" shows the number of items on which the taggers both identified the same error and offered the same or similar reconstructions. "X only" shows the number of cases in which tagger "X" identified an error that the other tagger in that pair did not.

Table 2.
Results of Experiment C

Essays	A&B				A&C				B&C		
	Same	A only	B only		Same	A only	B only		Same	B only	A only
1-10	305	169	137		305	167	145		288	123	156
11-22	139	67	43		144	42	122		163	45	108

One striking feature of the data in Table 2 is the drastic decrease in the number of errors tagged: the taggers identified approximately half as many errors in the second set of essays tagged as in the first set, indicating that one effect of the discussion between the two tagging sessions resulted in less concern with minor stylistic on the second set than

on the first.

A second, and somewhat disappointing, feature of the results in Table 2 is that the rate of agreement does not seem to have increased noticeably in the second set tagged as compared with the first. We will next explore the question of how to measure agreement between taggers.

Three measures of agreement between taggers have been used in the field:

1. **Recall:** How well does the performance of the "non-expert" match that of the "expert"?
2. **Precision:** What percentage of the "non-expert's" tags are accurate?
3. **Reliability:** How frequently do two taggers identify the same error?

Recall and precision were devised to measure the performance of a computer against that of a human "expert." A crucial aspect of these two measures is the designation of one member of the pair being compared as the "expert." Since this does not apply in any obvious way to our taggers -- all three had comparable training and experience, and all three underwent identical orientation to the task -- we arbitrarily designated the tagger who identified more errors as the expert. This was done as much for ease of calculation as for any principle. Reliability is a standard measure used in the social sciences to compare the performance of two taggers working on the same data, where one may or may not be considered an "expert."

We should note that none of these measures is entirely satisfactory. As noted, precision and recall assume that one of the two taggers being compared is an expert, which is not the case with our taggers. Reliability, while not assuming greater expertise on the part of one tagger, nevertheless fails to take into account the difficulty of the task. If there are only two possibilities for tagging an item -- true and false or succeed and fail -- an agreement rate of 50% is the level that could be achieved by chance. On the other hand, if there are 100 possible tags, an agreement rate of 50% might well be significant. However, the measure that Carletta (1996) proposes, the kappa-statistic, is not directly applicable to our data because it assumes a fixed number of possible tags. Our taggers have no such limitation, and therefore it is not clear how the kappa-statistic would apply. We merely note these difficulties in passing. We have no solution to the problem, so we have adopted the measures that are widely used in the field.

Tables 3, 4, and 5 present the results of the agreement scores for the three pairs of taggers.

Table 3. Tagging Results for Taggers A and B

Essay	A marked	B same	B marked	Recall	Precision	Reliability
1	56	45	64	0.8	0.7	0.5
2	31	12	28	0.39	0.43	-0.19
3	40	29	42	0.73	0.69	0.41
4	45	31	49	0.69	0.63	0.32
5	31	14	20	0.45	0.7	0.1
6	90	50	66	0.56	0.76	0.28
7	61	46	78	0.75	0.59	0.32
8	68	47	61	0.69	0.77	0.46
9	40	24	26	0.6	0.92	0.45
10	12	7	8	0.58	0.88	0.4
Total	474	305	442	0.64	0.69	0.33
Essay						
11	20	11	15	0.55	0.73	0.26
12	14	5	6	0.36	0.83	0
13	14	12	16	0.86	0.75	0.6
14	16	9	11	0.56	0.82	0.33
15	21	14	19	0.67	0.74	0.4
16	24	16	16	0.67	1.00	0.6
17	12	10	17	0.83	0.59	0.38
18	2	1	2	0.5	0.5	0
19	9	6	8	0.67	0.75	0.41
20	12	5	9	0.42	0.56	-0.05
21	29	22	31	0.76	0.71	0.47
22	33	28	32	0.85	0.88	0.72
Total	206	139	182	0.67	0.76	0.43

Table 4. Tagging Results for Taggers A and C

Essay	A marked	C same	C marked	Recall	Precision	Reliability
1	56	41	55	0.73	0.75	0.48
2	31	18	32	0.58	0.56	0.14
3	111	24	36	0.22	0.67	-0.35
4	40	27	42	0.68	0.64	0.32
5	30	15	21	0.5	0.71	0.18
6	93	52	70	0.56	0.74	0.28
7	59	39	65	0.66	0.6	0.26
8	71	53	79	0.75	0.67	0.41
9	39	28	36	0.72	0.78	0.49
10	13	8	14	0.62	0.57	0.19
Total	543	305	450	0.56	0.68	0.23
Essay						
11	15	11	24	0.73	0.46	0.13
12	6	4	12	0.67	0.33	-0.11
13	18	15	17	0.83	0.88	0.71
14	12	9	22	0.75	0.41	0.06
15	18	13	27	0.72	0.48	0.16
16	16	15	36	0.94	0.42	0.15
17	17	14	18	0.82	0.78	0.6
18	2	1	5	0.5	0.2	-0.43
19	8	3	16	0.38	0.19	-0.5
20	9	7	19	0.78	0.37	0
21	32	22	31	0.69	0.71	0.4
22	33	30	39	0.91	0.77	0.67
Total	186	144	266	0.77	0.54	0.27

Table 5. Tagging Results for Taggers B and C

Essay	C marked	B same	B marked	Recall	Precision	Reliability
1	57	41	62	0.72	0.66	0.38
2	33	20	29	0.61	0.69	0.29
3	39	26	40	0.67	0.65	0.32
4	49	31	39	0.63	0.79	0.41
5	21	14	20	0.67	0.70	0.37
6	62	38	62	0.61	0.61	0.23
7	61	38	64	0.62	0.59	0.22
8	75	48	60	0.64	0.80	0.42
9	36	25	27	0.69	0.93	0.59
10	11	7	8	0.64	0.88	0.47
TOTAL	444	288	411	0.72	0.70	0.35
Essay						
11	23	13	17	0.57	0.76	0.30
12	12	9	14	0.75	0.64	0.38
13	18	12	16	0.67	0.75	0.41
14	24	12	18	0.50	0.67	0.14
15	31	19	23	0.61	0.83	0.41
16	35	21	24	0.60	0.88	0.42
17	18	11	12	0.61	0.92	0.47
18	5	1	2	0.20	0.50	-0.43
19	16	5	9	0.31	0.56	-0.20
20	20	9	12	0.45	0.75	0.13
21	31	22	29	0.71	0.76	0.47
22	38	29	32	0.76	0.91	0.66
TOTAL	271	163	208	0.60	0.78	0.36

Table 6 summarizes the agreement data from Tables 3 through 5.

Table 6. Summary of Agreement Measures

Essays	A&B			A&C			B&C		
	Recall	Prec.	Rel.	Recall	Prec.	Rel.	Recall	Prec.	Rel.
1-10	0.64	0.69	0.33	0.56	0.68	0.23	0.65	0.73	0.37
11-22	0.67	0.76	0.43	0.77	0.54	0.27	0.55	0.74	0.26

4. Conclusions.

There are three major concrete results that are evident in the data in Table 6. First, the taggers tagged many fewer errors in the second set than in the first set. We assume that this is the result of the lengthy meetings held between the first and second tagging sessions. Evidently these meetings made that taggers aware that they should be looking for real errors, rather than stylistic changes that might make the essay better.

The second result is the disappointing finding that the procedure we used did not result in an increase in the rates of agreement between taggers.

The third result is that our training procedures did not succeed in achieving an adequate rate of agreement. In fact, the increase in agreement between the first and second set of essays is small to non-existent.

We conclude from these findings that we will never be able to assign a single tagger to a set of essays; more than one tagger will always be necessary. Furthermore, we will need to adopt some procedure for resolving the inevitable disagreements between taggers. Two solutions suggest themselves. We might hold regular meetings of the taggers to discuss the discrepancies, and attempt to reach a consensus on the correct reconstruction. Alternatively, we might have a third person whose function will be to review all of the tagged essays and make authoritative decisions about which ones to accept. In either case, it seems that differences in tagging is a fact of life that we will have to deal with.

REFERENCES

- Bardovi-Harlig, K. and Bofman, T. (1989). "Attainment of syntactic and morphological accuracy by advanced language learners". *Studies in Second Language Acquisition* 11: 17-34.
- Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, D. and Reppen, R. (1998). "Comparing native and learner perspectives on English grammar: a study of complement clauses" in Granger, S. (ed.) 1998.
- Carletta, Jean. (1996). "Assessing agreement on classification tasks: the Kappa statistic." *Computational Linguistics*. 22: 249-254.
- Fitzpatrick, Eileen, and M.S. Seegmiller (2000). "Experimenting with error tagging." Paper presented at the Annual Conference on Corpus Linguistics and Language Teaching, Northern Arizona University, April 2000.
- Granger, S. (ed.) (1998). *Learner English on Computer*. London: Longman.
- Kepner, C. (1991). "An experiment in the relationship of types of written feedback to the development of second-language writing skills". *Modern Language Journal* 75: 305-313.
- Meunier, F. (1998). Computer tools for the analysis of learner corpora. in Granger, S. (ed.) 1998.
- Milton, J. (1998). "Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment," in Granger, S. (ed.) 1998.
- Milton, J. and Chowdhury, N. (1994) "Tagging the interlanguage of Chinese learners of English." In Flowerdew, L. and Tong, K.K. (eds.) *Entering Text*. The Hong Kong University of Science and Technology, Hong Kong, pp. 127-43.
- Polio, C. (1997). "Measures of linguistic accuracy in second language writing research". *Language Learning* 47: 101-143.
- Zhang, S. (1987). "Cognitive complexity and written production in English as a second language". *Language Learning* 37: 469-481.

KEY WORDS: corpus tagging, error tagging, language-learner corpora, second-language acquisition, ESOL.