

The Montclair Electronic Language Database Project¹

Eileen Fitzpatrick

M.S.Seegmiller

Montclair State University

Abstract

The Montclair Electronic Language Database (MELD) is an expanding collection of essays written by students of English as a second language. This paper describes the content and structure of the database and gives examples of database applications.

The essays in MELD consist of the timed and untimed writing of undergraduate ESL students, dated so that progress can be tracked over time. Demographic data is also collected for each student, including age, sex, L1 background, and prior experience with English.

The essays are continuously being tagged for errors in grammar and academic writing as determined by a group of annotators. The database currently consists of 44,477 words of tagged text and another 53,826 words of text ready to be tagged.

The database allows various analyses of student writing, from assessment of progress over time to relation of error type and L1 background.

1. Introduction

A corpus of the productions of language learners provides authentic language data that can be analyzed and sampled for language performance. As Granger (1998) argues, the large size of a corpus, the naturalness of the data, and its computerization yield advantages that complement data collected in controlled experiments.

Corpus data represents the kind of data that learners use naturally. In addition, the data is collected from many informants, giving it a broad empirical base that enables descriptions of learner language to be generalized. Because of the size of the data set, even infrequent features of learner language can be studied, as well as the avoidance of difficult features of the language. A carefully constructed corpus can provide representative samples covering the different variables affecting learner productions. The large size of a corpus also sets the stage for innovations in teaching methodology and curriculum development as students examine learner data and compare it to native speaker language. Most

significant, the automated analysis of language has the "power to uncover totally new facts about language" (Granger 1998:3).

Language learner corpus building has been well established for more than ten years. Pravec (2002) discusses nine projects in Belgium, England, Hong Kong, Hungary, Japan, Poland, and Sweden, all of which represent the productions of foreign language learners.² Many of these corpora are annotated, giving them additional research value. The annotations include information on part of speech, syntactic structure, semantic relations, and type of error.

These corpora provide models of language performance that can be used to test hypotheses about the process of second language (L2) acquisition, to design teaching materials for the L2 writer, to design a parser for L2 writing, and to check the L2 writer's grammar. (Milton and Chowdhury, 1994).

The language learning experience of a foreign language learner is normally different from that of a second language learner, the latter being immersed in the language and required to use it on a daily basis. Indeed, Nickel (1989:298) observes that the lack of a distinction between the foreign and second language learner has been partly responsible for the contradictory results, particularly with respect to transfer, in SLA research. However, to date there has been no effort to build corpora comparable to the aforementioned data on foreign-language learners that represent the language of learners of English as a *second* language. The Montclair Electronic Language Database (MELD), under development at Montclair State University in the USA, aims to fill that gap in our understanding of the performance of English language learners.

MELD differs from the cited corpora not only in its capture of *second* language data, but also in its method for annotating errors in the data, and in its goal of making the data publicly available for the building of resources and tools for language learners and for researchers in L2 acquisition. A publicly available corpus will enable analyses to be duplicated and results to be shared. A corpus is a large investment in time, money, and equipment and the lack of access to corpus data diminishes the advantages that these collections provide.

This paper provides an overview of the MELD corpus, the annotation it provides, a discussion of its error annotation goals and techniques, sample applications using MELD data, and future plans for the project.

2. MELD Overview

The MELD corpus currently consists of formal essays written by upper level students of English as a Second Language preparing for college work in the United States. The corpus currently contains 44,477 words of text annotated for error and another 53,826 words waiting to be annotated. We expect to add another 50,000 words each year; if a funding source is found, we will accelerate this pace.

Some of the essays are timed essays written in class; the rest are written at home at the students' own pace. Essays are either submitted electronically or transcribed from hand-written submissions. A record is kept as to how each essay was submitted and whether it was written in a timed or untimed situation. Timed essays are written in class in response to a general prompt such as "If you had a choice between traditional schooling and studying at home with a computer, which would you choose?" These writing tasks are given to each class on entering and exiting the course. Untimed essays are written outside of class in response to a question about a reading or topic discussed in class. Both the timed and the untimed essays vary widely in length.

Participating student authors sign a release form that permits us to enter their written work into the corpus throughout the semester. These students also complete a background form on native language, other languages, schooling, and extent and type of schooling in the target language, currently only English. The background data for each student is stored in a flat file that links to the essays by that student. The writing of 65 students is currently represented in the database. The L1 languages represented are Arabic, Bengali, Chinese (Mandarin and Taiwanese), Haitian Creole, Gujarati, Hindi, Malayalam, Polish, Spanish, and Vietnamese. Close to a quarter of the students are multilingual. A portion of the background data and text data is currently web accessible.³

MELD currently has a small set of tools to enable entry, viewing and manipulation of both the student author background data and the text data. The student authors fill out a form asking for 21 items of background data including gender, age, native and other languages, and venues and methods of learning English. We have developed a pop-up window tool to ensure accurate entry of these data. Another tool enables the user to view student background data and retrieve the essays written by that student.

The data itself can also be viewed with the errors replaced by reconstructions. We hope that by using this viewer to remove low-level errors, annotator reliability might improve on errors that are more difficult to tag. We also have a crude concordancer that enables errors plus reconstructions to be viewed in context.

3. Data Annotation

3.1 Error Annotation

An important feature of MELD is the annotation of errors. Assuming that the goal of L2 learning is mastery of L1 performance, the value of a corpus of L2 productions lies in its ability to allow us to measure the distance between a sample of L2 writing and a comparable L1 corpus. Such a comparison also permits research into patterns of difference. The MELD annotation system allows such comparison.

Many of the differences between L1 and L2 corpora can be observed by online comparison of the two. 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 and sentence length (Meunier). An L2 corpus, however, also differs from a comparable L1 corpus in the number and type of morphological, syntactic, semantic, and rhetorical errors it exhibits, and this difference cannot be observed automatically; it requires the L2 text to be manually tagged for errors. To enable the researcher to find patterns, the individual errors must be tagged as errors and classified as to error type.

Systems of error classification often use a predetermined list of error types (see, for example, the studies cited in Polio, 1997). The Hong Kong corpus (Milton and Chowdhury, 1994) and the PELCRA corpus at the University of Lodz, Poland, use such a predetermined tagset (see Pravec, 2002). The main advantage of a predetermined list of error types is that it guarantees a high degree of tagging consistency among the annotators. However, a list limits the errors recognized to those in the tagset. Our concern in using a tagset was that we would skew the construction of a model of L2 writing by using a list that is essentially already a model of L2 errors, allowing annotators to overlook errors not on the list. The use of a tagset also introduces the possibility that annotators will misclassify those errors that do not fit neatly into one of the tags on the list.

In place of a tagset, our annotators "reconstruct" the error to yield an acceptable English sentence. Each error is followed by a slash, and a minimal reconstruction of the error is written within curly brackets. Missing items and items to be deleted are represented by "0". Tags and reconstructions look like this:

1. school systems {is/are}
2. since children {0/are} usually inspired
3. becoming {a/0} good citizens

The advantages of reconstruction over tagging from a predetermined tagset are that reconstruction is faster than classification, there is no chance of

misclassifying, and less common errors are captured. An added benefit is that a reconstructed text does not pose the problems for syntactic parsers and part-of-speech taggers that texts with ungrammatical forms pose (though see section 3.3). We anticipate that a reconstructed text can be more easily parsed and tagged for part-of-speech information than the unreconstructed essays.

Reconstruction, however, has its own difficulties. Without a tagset, annotators can vary greatly in what they consider an error. The wide discretion given to annotators results in annotation differences that run the gamut from the correction of clearly grammatical errors to stylistic revisions of rhetorical preferences. Even in the case of strictly grammatical errors, different annotators may reconstruct differently. For example, the common error represented in (4) can be reconstructed as either (5) or (6), and the less predictable (7) as either (8) or (9).

4. the student need help
5. the {student/students} need help
6. the student {need/needs} help.
7. We can also look up for anything that we might choose to buy,
8. We can also {look up/search} for anything that we might choose to buy,
9. We can also look {up/0} for anything that we might choose to buy,

We handle such discrepancies by adjudicating the tags as a team. Each text is tagged by two annotators, who then meet with a third annotator to discuss and resolve differences. For examples like (4) and (7), multiple reconstructions are entered, although we are aware that cases like (7) have several possible reconstructions.

More difficult issues involve grammatical rules that have a non-local domain. One recurring example involves the use of articles in English. For instance, the sentence

10. The learning process may be slower for {the/0} students as well

is correct with or without the article before *students*. However, the use of *the* indicates that a particular group of students had been identified earlier in the essay, whereas the absence of *the* indicates that *students* is being used in the generic sense. We choose to mark errors at the paragraph level; since no students had been identified earlier in the paragraph, we marked (10) as containing an error.

Language that involves the imposition of a standard also present difficulties for error tagging, primarily because the line between casual writing and academic writing is often fuzzy. Because of this vagueness, we are developing a list and using a different tag (square brackets) to annotate writing that violates an academic standard. Examples (11)-(12) illustrate this issue.

11. they learn how to interact with the other [kids/children]
12. [But/However] it doesn't take long for one to fit in

The blurred line between grammatical and rhetorical errors presents the most difficult error tagging problem. It is difficult to categorize examples like (13) and (14) as ungrammatical, yet the error in (13) fails to capture the rhetorical contrast of sad and happy while the choice of the present tense in (14) fails to adhere to tense concord.

13. they felt sad to live far from them {and/but} also happy because
14. Maybe I would have a problem that no computer {can/could} solve

3.2 Annotation agreement

Consistency among annotators is crucial if the annotation is to be useful. However, the fuzzy nature of many L2 learner errors makes consistency a serious concern. We have conducted several experiments on tagging consistency both between the authors (Fitzpatrick and Seegmiller, 2000) and among a group of ESL teachers (Seegmiller and Fitzpatrick, 2002). The consistency measures we have used for these experiments included interrater reliability, precision, and recall.

Interrater reliability (Polio, 1997) measures the percentage of errors tagged by both annotators, which we calculate as 1 minus the number of cases where one tagger or the other, but not both⁴, tagged an error divided by an average of the total number of errors tagged:

$$\text{Reliability} = 1 - \frac{T1 \oplus T2}{(T1+T2)/2}$$

This is the most stringent measure possible since we are calculating consistency on actual errors identified in common, not on number of errors identified, and we are not working from a predetermined set of errors, making every word and punctuation mark a target for an error tag. It was clear to us that our initial experiments might yield very low numbers and we could only hope that some basis for greater agreement would come out of the experiments.

Precision and recall are measures commonly used in evaluations of machine performance against a human 'expert'. We use these measures because they enable us to compare the performance of one annotator against the other so that we can address problems attributable to a single annotator. To obtain these measures, we arbitrarily assume one annotator to be the expert.

Precision measures the percentage of the non-expert's tags that are accurate. It is represented as the intersection (\cap) of the non-expert's (T2) tags with the expert's tags (T1) divided by all of T2's tags.

$$\text{Precision} = T1 \cap T2 / T2$$

For example, if T1 tagged 25 errors in an essay and T2 tagged the same 25 errors but also tagged 25 more errors not tagged by T1, then T2's precision rate would be .5

Recall measures the percentage of true errors the non-expert found. It is represented as the intersection of the non-expert's (T2) tags with the expert's tags (T1) divided by all of T1's tags.

$$\text{Recall} = T1 \cap T2 / T1$$

Following our example above, T2's recall would be 1.0 since T2 tagged all the items that T1 tagged.

Precision and recall can be illustrated as in Figure 1, which shows one possible outcome of the performance of two annotators. The non-expert has achieved high precision in this task; most of the errors she tagged were identified by the expert as errors. However her recall rate is low; she missed about half of the errors identified by the expert.

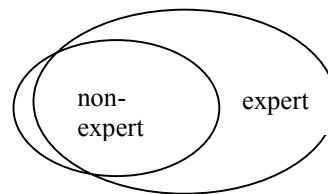


Figure 1. Precision and Recall for an expert and a non-expert tagger.

We might expect the situation represented in Figure 1 if there are many low level grammatical errors that both annotators tagged as well as another type of error, e.g., errors involving academic writing standards, that T1 tagged but T2 did not. The precision and recall measures allow us to track the overzealous tagger and discover the source of a pattern of tagging disagreements.

Both experiments that we conducted into tagger agreement involved two tests. The first test let the annotators tag errors with no instruction. This was followed by a meeting in which the taggers established general guidelines for tagging that then guided test two. Table 1 shows the results of these two tests with the authors as annotators. Tables 2-4 show the pair-wise results among three ESL teachers

who acted as taggers. The data sets were the same for both experiments; set one contained 2476 words, and set two 2418. The error counts indicated were those of the 'expert'; the teachers rotated as experts.

Table 1. Results with authors as annotators

Data set	Errors	Recall	Precision	Reliability
One	241	.73	.84	.54
Two	193	.76	.90	.60

Table 2. Results with J&L as annotators

Essay	Errors	Recall	Precision	Reliability
One	474	.54	.58	.39
Two	206	.57	.78	.49

Table 3. Results with J&N as annotators

Essay	Errors	Recall	Precision	Reliability
One	472	.58	.48	.23
Two	186	.37	.54	.27

Table 4. Results with L&N as annotators

Essay	Errors	Recall	Precision	Reliability
One	411	.65	.70	.37
Two	208	.60	.78	.36

These levels of agreement are clearly unsatisfactory, and have led to our present practice of resolving disagreements between annotators by adjudication with a third annotator. Unfortunately, this is expensive and slows the task considerably.

Since taggers differ in the extent to which they mark stylistic and rhetorical features of the essays, another helpful solution has been to use a different type of mark for errors involving a written standard, as mentioned in the previous section. These errors, particularly errors involving punctuation, verb mood (*if I [have/had] the chance*), and certain lexical choices (*the [kids/children] can*) make up a large proportion of the disagreements. It has proven effective to separate these from the language acquisition errors.

1.3 Part-of-Speech tagging

Since automatic part-of-speech (pos) tagging and parsing are built on models of grammatical English, we anticipated that reconstructing errors would aid in the application of these systems to our data. To date, one test of an automatic pos tagger, the Brill tagger (Brill, 1995), has assessed the performance of an automatic system on a test set of both the uncorrected and corrected MELD data (Higgins, 2002).

The pos test included six essays, involving 1521 words of raw text and 1551 words of reconstructed text. Once difficulties with contractions and parentheses were removed, only 22 errors appeared in both sets of essays, an additional four appeared in the raw text alone and another two in the reconstructed text. This gives an error rate of .017 percent on the raw text and .015 on the reconstructed text. We assume that the high accuracy of the Brill tagger, even on the raw data, resulted from the highly proficient writing currently represented in MELD. We still assume that as we capture the writing of less proficient learners, the reconstruction of errors will aid the pos tagging.

The small number of pos tagging errors indicates that automatic pos tagging is a reasonable enhancement to the MELD data. Equally encouraging is the fact that the most common pos tagging error, with 10 occurrences, was caused by the labeling of capitalized ordinal numbers as proper nouns by the Brill tagger.⁵

4. Possible Applications

MELD, at under 100,000 words, is still a small corpus. However, even with a small corpus, there are trends that we can observe, particularly if we look at the raw data. Looking at the smaller, tagged portion of the corpus, we can present research that is illustrative of what can be done with a tagged corpus.

4.1 Studies of progress over a semester⁶

Since the data in MELD include longitudinal data in the form of essays written by the same student over the course of a semester or more, one of the possible applications of the data is the study of changes in student writing over time. In this section, we will present some examples of the study of such changes using both the untagged and the tagged versions of the essays.

When assessing students' writing over time, there are certain changes that we expect to find if our English-language program is working effectively. If we compare a timed essay written at the beginning of the semester with one written at the end, we would expect to find, among others, the following sorts of changes:

1. Fluency will increase. That is, students will be able to write more easily, without having to stop and think about what to say and how to say it.

2. Sentences will get longer. As students' command of the target language increases, they will become more confident in their use of longer sentences.

3. Sentences will become more complex. A sentence can be long but fairly simple, for example if it consists of several simple clauses joined by conjunctions ("John got up and he took a shower and he shaved and he got dressed"). But it is an indication of increasing mastery of the syntax of the L2 when students begin to use more complex sentence types ("After getting up but before getting dressed, John showered and shaved"). Sentence complexity is notoriously difficult to measure and many different approaches have been proposed, but one simple one is to count the number of clauses per sentence, measured by counting the number of verbs.

4. Vocabulary will increase. It is difficult to measure any person's total vocabulary. One approach is to count the number of different words used in a timed essay and take that as a rough measure of overall vocabulary. This approach assumes that students with a limited vocabulary will tend to use the same words over and over again, whereas students with a greater command of the language will be able to use a greater variety of words in an essay of limited length.

5. The number of errors will decrease. For obvious reasons, this is usually taken as a standard measure of mastery of a language.

In this illustrative study, we used two essays from each of 23 students, for a total of 46 essays. One essay was written at the beginning of the semester and the other at the end, allowing us to measure what kinds of changes occur in the students writing during a rigorous ESOL writing course. The essays vary greatly in length, ranging from 86 to 377 words. The authors of the essays are from a variety of L1 backgrounds. Our analyses made use of several standard UNIX text-processing tools, although similar studies could be carried out with any of several software packages. It should be noted that since the results reported below are for purposes of illustration, we have not carried out any statistical calculations to determine which, if any, results are statistically significant.

For our first study, we calculated the mean length of the essays and compared the essays written at the beginning of the semester (the Pre-Test) with those written at the end of the semester (the Post-Test). The results are shown on Table 5.

Table 5. Mean Number of Words Per Essay

Pre-Test	Post-Test
189.1	236.8

As anticipated, the average number of words per essay increased (substantially, in fact), indicating that in a 20-minute timed essay, students were able to write much more at the end of the semester than they were at the beginning.

Next, we counted mean sentence length, which provides a rough but easy measure of syntactic complexity. Table 6 shows the results for 23 students:

Table 6. Mean Sentence Length

	Pre-Test	Post-Test
Mean words per sentence	18.2	18.8

While this is a far less dramatic result, we still get a change in the predicted direction: the number of words per sentence has increased.

Next we looked at a slightly more sophisticated measure of sentence complexity, the number of clauses per sentence, measured by simply counting the number of main verbs per clause and dividing by the number of sentences:

Table 7. Number of Clauses Per Sentence

	Pre-Test	Post Test
Mean clauses per sentence	3.6	3.4

There is actually a slight decrease in the number of clauses per sentence, a phenomenon that might deserve further investigation. The next logical step in investigating changes in sentence complexity would be to separate conjoined clauses from embedded clauses, since the latter are more complex. It is possible that the students are using fewer but more complex clauses, or perhaps they have simply learned that shorter sentences are more effective.

Next we looked at changes in the number of errors in the essays. Errors are easy to count in the tagged essays. Here are the data:

Table 8. Errors

	Pre-Test	Post-Test
Mean number of errors/sentence	1.46	1.34

Once again, we find the expected result: a decrease in the number of errors per sentence.

Finally, we counted the number of different words used in each essay and then calculated the type/token ratio to control for the differing lengths of the essays. Table 9 shows the vocabulary results:

Table 9. Vocabulary

	Pre-Test	Post-Test
Mean total words	189.1	224.5
Mean vocabulary	102.7	117.5
Mean Type/Token Ratio	1.81	1.91

We see that both the total vocabulary and the type token ratio increase between the first and second essays.

Incidentally, when UNIX is used, one of the byproducts of the measure of vocabulary is a word frequency count, which is a list of all the words in a text with the frequency of each, arranged from most to least frequent. This is an interesting document in its own right, and might be studied in various ways, for example to see how many unusual (as opposed to common) words a student uses. In one of our studies, it was noticed that the relative frequency of *the* was about the same for speakers of Spanish as for those of English – *the* is typically the first or second most common word in the text – while for speakers of Japanese and Russian, *the* occurred much less frequently, often ranking as low as the fifteenth most frequent word.

4.2 Research on error types by L1

Concordancing of the errors currently enables us to compare problematic points with points students have mastered. For example, the essays so far demonstrate difficulty in mastering the correct preposition in a prepositional phrase complement to a head noun, verb or adjective when there is some notion of movement involved, as the following data, with the head in boldface, show.

even before he **paid** {to/0} his Aunt Mary for his
 have when they **arrive** {to/in} the new country Entering into
 four months of his **arrival** {to/in} the United States so he
 Mike and Mary **departing** {to/for} the United States In
 were sad to **separate** {to/from} their son and daughter, but
 had never been **separated** {to/from} their family before Mr

see the parents **separated** {to/from} them {0/} they felt sad
 at the time they **separated** {to/from} them they felt sad
 are closing the **door** {from/to} Ireland I {0/am} going to write
 sad to live **far** {of/from} them [and/but] also happy because

In contrast, the same data shows a good command of prepositional complements to abstract nouns and verbs:

they are taking the **risk** of hurting their parent's feelings
 time they also have **fear** of taking risks in unknown
 country may be **thought** of as opening a door
 never have the same **relation** with their siblings They would
 as family and their **relationship** to their land and country
 may never loose their **relationship** with their family and friends
 her life They were **dreaming** about their future but they
 happy life she always **dreamed** of. As mentioned in the

Coupled with the demographic information, the error tagging also permits the correlation of grammatical properties with speaker differences. For example, in our still small data set, we see the following errors in concord between tenses for native speakers of Spanish and Gujarati.

Spanish:

Mr. and Mrs. Feeney worked their whole lives to {gave/give} a good education
 they understood that Mary and Michael {can/could} have a better future
 The risks they take occur when they {went/go} to the United States

Gujarati:

he would send money home when he {would start/started} earning it
 might not be able to see to help that person whenever they {will/0} {need/needs}
 If Michael and Mary {will be/are} successful
 they are not {use/used} to different types of work
 they got used to it as time {passes/passed} by.
 They were dreaming, but they {do/did} not know what {is/was} going to happen.

However, in a 2300 word sample produced by four native Spanish speakers there were only 6 such errors, while in the same size sample produced by four Gujarati speakers there were 31 such errors, a five-fold difference in mastery of tense concord. While the samples are small, the difference is striking and the grammatical phenomenon -- concord between tenses -- would probably go unnoticed without the systematic view of the data given by the corpus.

4.3 Preparation of instructional materials

The type and frequency of error by level or native language background guides the teacher to writing problems that students of a comparable level and background need to work on. One can also compare the corpus to the work of comparable, proficient native writers to discover gaps in the L2 writing and develop materials accordingly. The corpus can also be used for testing purposes since it allows testing to be targeted to specific levels and language backgrounds.

Several types of corpus-based exercises for students have been developed (e.g., Milton, 1998) though they are not widely available. A publicly available corpus will enable more exercises of this type. Students can also use portions of the corpus for proof reading exercises, with the reconstructed text available for checking. Certain types of error can be 'turned off' so that the student sees only the type of usage s/he needs to master. The student can then compare corrections with those of the annotator.

5. Conclusion

MELD is a small but growing database of learner writing. It is accessible on line to anyone who wishes to use it,³ and the tools for searching and analyzing the data will continue to be expanded. We also hope to add data from other institutions, as well as spoken data from L2 learners.

Along with the gradual increase in tagged data, we plan to enhance access to MELD and build tools that will enhance the usefulness of the data. We anticipate bringing certain tools online in the near future; some tool development requires funding that puts it beyond our immediate capability.

Among our immediate goals are improved online access to the data, including the use of a concordancer to view errors and reconstructions, automatic part of speech annotation as a user option, and the addition of data from different ESL skill levels. Our long range plans include a statistical tool to correlate error frequency with student background; student editing aids, most specifically a grammar checker using our current data as a model; and -- dream of dreams -- the addition of L2 spoken data.

The data in MELD can be used for a variety of both research and educational purposes, including the study of L2 acquisition and the preparation of teaching materials. It is our hope that MELD will prove to be a valuable resource to our colleagues in the field of second-language acquisition and teaching.

Notes

¹ We wish to thank the master teachers Jacqueline Cassidy, Norma Pravec, and Lenore Rosenbluth, who contributed careful labor and thoughtful discussion in providing a tagged data set and tagging guidelines and the graduate student annotators Jennifer Higgins, Donna Samko, and Jory Samkoff, and the programmers and data entry personnel Jennifer Higgins and Kae Shigeta.

² The corpora created in England (the Cambridge Learner Corpus and the Longman Corpus) represent the writing of students in non-English-speaking countries.

³ Student background data and essays are available at <http://www.chss.montclair.edu/linguistics/MELD>.

⁴ \oplus is to be interpreted as 'exclusive or', indicating that if one tagger marked a feature as an error, the other tagger did not.

⁵ The Brill tags are based on the manually tagged labels of the Penn Treebank (Marcus 1993), which labels all the items in a name like *First National City Bank* as proper nouns, giving *First*, *Second*, etc. a high frequency as a proper noun.

⁶ Some of the material in this section was presented in Seegmiller et al (1999).

References

- Biber, D. and R. Reppen (1998), 'Comparing native and learner perspectives on English grammar: a study of complement clauses', in: Granger, S. (ed.) *Learner English on Computer*, London: Longman, 145-158.
- Brill, E. (1995), 'Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging'. *Computational Linguistics*, 21(4), 543-566.
- Fitzpatrick, E. and M.S. Seegmiller (2000), 'Experimenting with Error Tagging in a language learning corpus'. *The Second North American Symposium of the American Association for Applied Corpus Linguistics*. Northern Arizona University, Flagstaff, AZ, March 31-April 2.
- Granger, S. (ed.) (1998), *Learner English on Computer*. London: Longman.
- Granger, S. (1998), 'The computer learner corpus: a versatile new source of data for SLA research', in: Granger, S. (ed.) *Learner English on Computer*, London: Longman, 3-18.
- Higgins, J. (2002), 'Comparing the Performance of the Brill Tagger on Corrected and Uncorrected Essays', <http://picard.montclair.edu/linguistics/MELD/pos.html>.
- Marcus, M., Santorini, B., and Marcinkiewicz, M. (1993), 'Building a large annotated corpus of English: the Penn Treebank'. *Computational Linguistics*, 19(2), 313-330.

- Meunier, F. (1998), 'Computer tools for the analysis of learner corpora', in: Granger, S. (ed.), *Learner English on Computer*, London: Longman, 19-38.
- Milton, J. (1998), 'Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment', in: Granger, S. (ed.) *Learner English on Computer*, London: Longman, 186-198.
- Milton, J. and N. Chowdhury (1994), 'Tagging the interlanguage of Chinese learners of English', in L. Flowerdew and A.K.K. Tong (eds.) *Entering Text*, Language Centre, The Hong Kong University of Science and Technology.
- Nickel, G. (1989), 'Some controversies in present-day error analysis: "contrastive" vs. "non-contrastive" errors', *International Review of Applied Linguistics* 27:292-305.
- Polio, C. (1997), 'Measures of Linguistic Accuracy in Second Language Writing Research', *Language Learning*, 47: 101-143.
- Pravec, N. (2002), 'Survey of learner corpora', *ICAME Journal*, 26: 81-114.
- Seegmiller, M.S., E. Fitzpatrick, and M. Call (1999), 'Assessing Language Development: Using Text-Processing Tools in Second-Language Teaching and Research', MEXTESOL, Mazatlan, MX.
- Seegmiller, M.S. and E. Fitzpatrick (2002), 'Practical Aspects of Corpus Tagging', in B. Lewandowska-Tomaszczyk and P.J. Melia (eds.) *PALC '01: Practical Applications in Language Corpora*. New York: Peter Lang.