A Corpus of Textual Revisions in Second Language Writing

John Lee and Jonathan Webster
The Halliday Centre for Intelligent Applications of Language Studies
Department of Chinese, Translation and Linguistics
City University of Hong Kong
{jsylee,ctjjw}@cityu.edu.hk

Abstract

This paper describes the creation of the first large-scale corpus containing drafts and final versions of essays written by non-native speakers, with the sentences aligned across different versions. Furthermore, the sentences in the drafts are annotated with comments from teachers. The corpus is intended to support research on textual revision by language learners, and how it is influenced by feedback. This corpus has been converted into an XML format conforming to the standards of the Text Encoding Initiative (TEI).

1 Introduction

Learner corpora have been playing an increasingly important role in both Second Language Acquisition and Foreign Language Teaching research (Granger, 2004; Nesi et al., 2004). These corpora contain texts written by non-native speakers of the language (Granger et al., 2009); many also annotate text segments where there are errors, and the corresponding error categories (Nagata et al., 2011). In addition, some learner corpora contain pairs of sentences: a sentence written by a learner of English as a second language (ESL), paired with its correct version produced by a native speaker (Dahlmeier and Ng, 2011). These datasets are intended to support the training of automatic text correction systems (Dale and Kilgarriff, 2011).

Less attention has been paid to how a language learner produces a text. Writing is often an iterative and interactive process, with cycles of textual revision, guided by comments from language teachers.

Discipline	# drafts
Applied Physics	988
Asian and International Studies	410
Biology	2310
Building Science and Technology	705
Business	1754
Computer Science	466
Creative Media	118
Electronic Engineering	1532
General Education	651
Law	31
Linguistics	2165
Management Sciences	1278
Social Studies	912
Total	13320

Table 1: Draft essays are collected from courses in various disciplines at City University of Hong Kong. These drafts include lab reports, data analysis, argumentative essays, and article summaries. There are 3760 distinct essays, most of which consist of two to four successive drafts. Each draft has on average 44.2 sentences, and the average length of a sentence is 13.3 words. In total, the corpus contains 7.9 million words.

Understanding the dynamics of this process would benefit not only language teachers, but also the design of writing assistance tools that provide automatic feedback (Burstein and Chodorow, 2004).

This paper presents the first large-scale corpus that will enable research in this direction. After a review of previous work (§2), we describe the design and a preliminary analysis of our corpus (§3).

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 248–252,
Jeju, Republic of Korea, 8-14 July 2012.

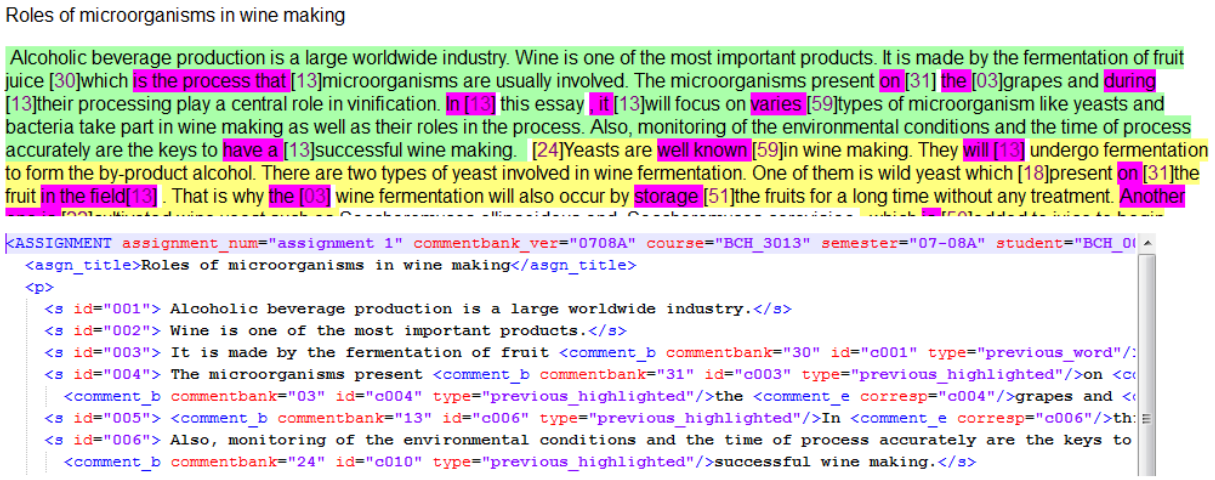

Figure 1: On top is a typical draft essay, interleaved with comments from a tutor (§3.2): two-digit codes from the Comment Bank are enclosed in angled brackets, while open-ended comments are enclosed in angled brackets. On the bottom is the same essay in TEI format, the output of the process described in §3.3.

2 Previous Research

In this section, we summarize previous research on feedback in language teaching, and on the nature of the revision process by language learners.

2.1 Feedback in Language Learning

Receiving feedback is a crucial element in language learning. While most agree that both the *form* and *content* of feedback plays an important role, there is no consensus on their effects. Regarding form, some argue that direct feedback (providing corrections) are more effective in improving the quality of writing than indirect feedback (pointing out an error but not providing corrections) (Sugita, 2006), but others reached opposite conclusions (Ferris, 2006; Lee, 2008).

Regarding content, it has been observed that teachers spend a disproportionate amount of time on identifying word-level errors, at the expense of those at higher levels, such as coherence (Furneaux et al., 2007; Zamel, 1985). There has been no large-scale empirical study, however, on the effectiveness of feedback at the paragraph or discourse levels.

2.2 Revision Process

While text editing in general has been analyzed (Mahlow and Piotrowski, 2008), the nature of revisions by language learners — for example, whether learners mostly focus on correcting mechanical, word-level errors, or also substantially reorganize paragraph or essay structures — has hardly been investigated. One reason for this gap in the literature is the lack of corpus data: none of the existing learner corpora (Izumi et al., 2004; Granger et al., 2009; Nagata et al., 2011; Dahlmeier and Ng, 2011) contains drafts written by non-native speakers that led to the "final version". Recently, two corpora with text revision information have been compiled (Xue and Hwa, 2010; Mizumoto et al., 2011), but neither contain feedback from language teachers. Our corpus will allow researchers to not only examine the revision process, but also investigate any correlation with the amount and type of feedback.

3 Corpus Description

We first introduce the context in which our data was collected (§3.1), then describe the kinds of comments in the drafts (§3.2). We then outline the conversion process of the corpus into XML format (§3.3), followed by an evaluation (§3.4) and an analysis (§3.5).

3.1 Background

Between 2007 and 2010, City University of Hong Kong hosted a language learning project where English-language tutors reviewed and provided feedback on academic essays written by students,

Paragraph level		Sentence level		Word level	
Coherence: more elaboration is needed	680	Conjunction missing	1554	Article missing	10586
Paragraph: new paragraph	522	Sentence: new sentence	1389	Delete this	9224
Coherence: sign posting	322	Conjunction: wrong use	923	Noun: countable	7316
Coherence: missing topic sentence	222	Sentence: fragment	775	Subject-verb agreement	4008

Table 2: The most frequent error categories from the Comment Bank, aimed at errors at different levels.

most of whom were native speakers of Chinese (Webster et al., 2011). More than 300 TESOL students served as language tutors, and over 4,200 students from a wide range of disciplines (see Table 1) took part in the project.

For each essay, a student posted a first draft[1] as a blog on an e-learning environment called Blackboard Academic Suite; a language tutor then directly added comments on the blog. Figure 1 shows an example of such a draft. The student then revised his or her draft and may re-post it to receive further comments. Most essays underwent two revision cycles before the student submitted the final version.

3.2 Comments

Comments in the draft can take one of three forms:

Code The tutor may insert a two-digit code, representing one of the 60 common error categories in our "Comment Bank", adopted from the XWiLL project (Wible et al., 2001). These categories address issues ranging from the word level to paragraph level (see Table 2), with a mix of direct (e.g., "new paragraph") and indirect feedback (e.g., "more elaboration is needed").

Open-ended comment The tutor may also provide personally tailored comments.

Hybrid Both a code and an open-ended comment.

For every comment[2], the tutor highlights the problematic words or sentences at which it is aimed. Sometimes, general comments about the draft as a whole are also inserted at the beginning or the end.

[1] In the rest of the paper, these drafts will be referred to "version 1", "version 2", and so on.

[2] Except those comments indicating that a word is missing.

3.3 Conversion to XML Format

The data format for the essays and comments was not originally conceived for computational analysis. The drafts, downloaded from the blog entries, are in HTML format, with comments interspersed in them; the final versions are Microsoft Word documents. Our first task, therefore, is to convert them into a machine-actionable, XML format conforming to the standards of the Text Encoding Initiative (TEI). This conversion consists of the following steps:

Comment extraction After repairing irregularities in the HTML tags, we eliminated attributes that are irrelevant to comment extraction, such as font and style. We then identified the Comment Bank codes and open-ended comments.

Comment-to-text alignment Each comment is aimed at a particular text segment. The text segment is usually indicated by highlighting the relevant words or changing their background color. After consolidating the tags for highlighting and colors, our algorithm looks for the nearest, preceding text segment with a color different from that of the comment.

Title and metadata extraction From the top of the essay, our algorithm scans for short lines with metadata such as the student and tutor IDs, semester and course codes, and assignment and version numbers. The first sentence in the essay proper is taken to be the title.

Sentence segmentation Off-the-shelf sentence segmentators tend to be trained on newswire texts (Reynar and Ratnaparkhi, 1997), which significantly differ from the noisy text in our corpus. We found it adequate to use a stop-list, supplemented with a few regular expressions

Evaluation	Precision	Recall
Comment extraction		
- *code*	94.7%	100%
- *open-ended*	61.8%	78.3%
Comment-to-text alignment	86.0%	85.2%
Sentence segmentation	94.8%	91.3%

Table 3: Evaluation results of the conversion process described in §3.3. Precision and recall are calculated on correct detection of the start and end points of comments and boundaries.

that detect exceptions, such as abbreviations and digits.

Sentence alignment Sentences in consecutive versions of an essay are aligned using cosine similarity score. To allow dynamic programming, alignments are limited to one-to-one, one-to-two, two-to-one, or two-to-two[3]. Below a certain threshold[4], a sentence is no longer aligned, but is rather considered inserted or deleted. The alignment results are stored in the XCES format (Ide et al., 2002).

3.4 Conversion Evaluation

To evaluate the performance of the conversion algorithm described in §3.3, we asked a human to manually construct the TEI XML files for 14 pairs of draft versions. These gold files are then compared to the output of our algorithm. The results are shown in Table 3.

In comment extraction, codes can be reliably identified. Among the open-ended comments, however, those at the beginning and end of the drafts severely affected the precision, since they are often not quoted in brackets and are therefore indistinguishable from the text proper. In comment-to-text alignment, most errors were caused by inconsistent or missing highlighting and background colors.

The accuracy of sentence alignment is 89.8%, measured from the perspective of sentences in Version 1. It is sometimes difficult to decide whether a sentence has simply been edited (and should therefore be aligned), or has been deleted with a new sentence inserted in the next draft.

[3]That is, the order of two sentences is flipped.

[4]Tuned to 0.5 based on a random subset of sentence pairs.

3.5 Preliminary Analysis

As shown in Table 4, the tutors were much more likely to use codes than to provide open-ended comments. Among the codes, they overwhelmingly emphasized word-level issues, echoing previous findings (§2.1). Table 2 lists the most frequent codes. Missing articles, noun number and subject-verb agreement round out the top errors at the word level, similar to the trend for Japanese speakers (Lee and Seneff, 2008). At the sentence level, conjunctions turn out to be challenging; at the paragraph level, paragraph organization, sign posting, and topic sentence receive the most comments.

In a first attempt to gauge the utility of the comments, we measured their density across versions. Among Version 1 drafts, a code appears on average every 40.8 words, while an open-ended comment appears every 84.7 words. The respective figures for Version 2 drafts are 65.9 words and 105.0 words. The lowered densities suggest that students were able to improve the quality of their writing after receiving feedback.

Comment Form	Frequency
Open-ended	47072
Hybrid	1993
Code	88370
- *Paragraph level*	*3.2%*
- *Sentence level*	*6.0%*
- *Word level*	*90.8%*

Table 4: Distribution of the three kinds of comments (§3.2), with the Comment Bank codes further subdivided into different levels (See Table 2).

4 Conclusion and Future Work

We have presented the first large-scale learner corpus which contains not only texts written by non-native speakers, but also the successive drafts leading to the final essay, as well as teachers' comments on the drafts. The corpus has been converted into an XML format conforming to TEI standards.

We plan to port the corpus to a platform for text visualization and search, and release it to the research community. It is expected to support studies on textual revision of language learners, and the effects of different types of feedback.

Acknowledgments

We thank Shun-shing Tsang for his assistance with implementing the conversion and performing the evaluation. This project was partially funded by a Strategic Research Grant (#7008065) from City University of Hong Kong.

References

Jill Burstein and Martin Chodorow. 2004. Automated Essay Evaluation: The Criterion online writing service. *AI Magazine*.

Daniel Dahlmeier and Hwee Tou Ng. 2011. Grammatical Error Correction with Alternating Structure Optimization. *Proc. ACL.*

Robert Dale and Adam Kilgarriff. 2011. Helping Our Own: The HOO 2011 Pilot Shared Task. *Proc. European Workshop on Natural Language Generation (ENLG)*, Nancy, France.

Dana Ferris. 2006. Does Error Feedback Help Student Writers? New Evidence on the Short- and Long-Term Effects of Written Error Correction. In *Feedback in Second Language Writing: Contexts and Issues*, Ken Hyland and Fiona Hyland (eds). Cambridge University Press.

Clare Furneaux, Amos Paran, and Beverly Fairfax. 2007. Teacher Stance as Reflected in Feedback on Student Writing: An Empirical Study of Secondary School Teachers in Five Countries. *International Review of Applied Linguistics in Language Teaching* 45(1): 69-94.

Sylviane Granger. 2004. Computer Learner Corpus Research: Current Status and Future Prospect. *Language and Computers* 23:123–145.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. International Corpus of Learner English v2. Presses universitaires de Louvain, Belgium.

Nancy Ide, Patrice Bonhomme, and Laurent Romary. 2000. XCES: An XML-based Encoding Standard for Linguistic Corpora. *Proc. LREC*.

Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. The NICT JLE Corpus: Exploiting the Language Learners' Speech Database for Research and Education. *International Journal of the Computer, the Internet and Management* 12(2):119–125.

Icy Lee. 2008. Student Reactions to Teacher Feedback in Two Hong Kong Secondary Classrooms. *Journal of Second Language Writing* 17(3):144-164.

John Lee and Stephanie Seneff. 2008. An Analysis of Grammatical Errors in Nonnative Speech in English. *Proc. IEEE Workshop on Spoken Language Technology.*

Cerstin Mahlow and Michael Piotrowski. 2008. Linguistic Support for Revising and Editing. *Proc. International Conference on Computational Linguistics and Intelligent Text Processing*.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. *Proc. IJCNLP*.

Ryo Nagata, Edward Whittaker, and Vera Sheinman. 2011. Creating a Manually Error-tagged and Shallow-parsed Learner Corpus. *Proc. ACL.*

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A Maximum Entropy Approach to Identifying Sentence Boundaries. *Proc. 5th Conference on Applied Natural Language Processing*, Washington DC.

Yoshihito Sugita. 2006. The Impact of Teachers' Comment Types on Students' Revision. *ELT Journal* 60(1):34–41.

Hilary Nesi, Gerard Sharpling, and Lisa Ganobcsik-Williams. 2004. Student Papers Across the Curriculum: Designing and Developing a Corpus of British Student Writing. *Computers and Composition* 21(4):439–450.

Frank Tuzi. 2004. The Impact of E-Feedback on the Revisions of L2 Writers in an Academic Writing Course. *Computers and Composition* 21(2):217-235.

Jonathan Webster, Angela Chan, and John Lee. 2011. Online Language Learning for Addressing Hong Kong Tertiary Students' Needs in Academic Writing. *Asia Pacific World* 2(2):44–65.

David Wible, Chin-Hwa Kuo, Feng-Li Chien, Anne Liu, and Nai-Lung Tsao. 2001. A Web-Based EFL Writing Environment: Integrating Information for Learners, Teachers, and Researchers. *Computers and Education* 37(34):297-315.

Huichao Xue and Rebecca Hwa. 2010. Syntax-Driven Machine Translation as a Model of ESL Revision. *Proc. COLING*.

Vivian Zamel. 1985. Responding to Student Writing. *TESOL Quarterly* 19(1):79-101.