MUCH: THE MALMÖ UNIVERSITY-CHALMERS CORPUS OF ACADEMIC WRITING AS A PROCESS

ANDREAS ERIKSSON[1]; DAMIAN FINNEGAN[2], ASKO KAUPPINEN[2], MARIA WIKTORSSON[2], ANNA WÄRNSBY[2]; PETER WITHERS[3]
[1]Chalmers University of Technology, Gothenburg
[2]Malmö University, Malmö
[3] Max Planck Institute for Psycholinguistics, Nijmegen
andreas.eriksson@chalmers.se, anna.warnsby@mah.se

Introduction

Recent development in European higher education shows an increasing need for and interest in writing research and writing pedagogy. This development is evidenced in the establishment of associations such as the European Association for the Teaching of Academic Writing (EATAW) and their publication Journal of Academic Writing. Another sign of the growing importance of writing research and pedagogy is the establishment of various types of writing centres and writing units at universities in Europe based on a model from American universities: University of Coventry (2004); European University Viadrina (2007); Chalmers University of Technology (2008); University of Copenhagen (2008); Stockholm University (2007), and Malmö University (2011). The activity and mission of these centres vary significantly, but their formation clearly exemplifies the increasing interest in academic writing within Europe.

The use of corpora in relation to academic writing has so far seen two major L1-based projects: the British Academic Written English corpus (BAWE) and Michigan Corpus of Upper-Level Student Papers (MICUSP), as well as a great number of projects based on small-scale written corpora (e.g. Gavioli 2005, see Boulton 2010 for an overview of projects). At present there are several corpus projects that aim at building non-native academic writing corpora, for example the Lancaster Corpus of Academic Written English (LANCAWE), the Corpus of Academic Learner English (CALE), and The Varieties of English for Specific Purposes dAtabase (VESPA) learner corpus.

The contribution of corpus-based research to writing research and writing pedagogy specifically has so far primarily involved aspects of vocabulary (Coxhead 2000), multi-word units or lexical bundles (Hyland 2008), grammar (Hinkel 2004), and approaches to teaching these features, either via so-called data-driven learning (DDL) (Boulton 2009) or a corpus-based approach. The tendency of corpus-based projects to focus on the linguistic aspects of academic writing is also evidenced in Flowerdew's (2010) comprehensive overview of how corpora have been used in writing instruction. There are obviously notable exceptions to this tendency (see e.g. Charles 2007 and Flowerdew 2008), but it still seems that it should be possible to broaden the scope of what a learner corpus is and can be used for in connection to research on academic writing.

In addition, up until now, systematically compiled corpora of written learner language have almost exclusively focused on a single (typically a final) version of a text. This focus obviously has certain advantages, for instance in terms of showing the students' actual performance, but it does not reflect how most texts are produced. Berkenkotter & Huckin addressed the problem of investigating merely the final version of a text when they argued

that: “Although something can be gained by studying published reports, certainly, we feel that tracking and analyzing the development of a report as it goes through various revisions yields unique insights about the epistemology of science” (1995:49). Because corpora have not been compiled for the purpose of analyzing writing processes, they can say very little about the development of a text. In order to address writing from a wider variety of perspectives, systematically collected material that is easily available and that covers several stages of the writing process are needed.

What sets the MUCH-corpus apart from other learner corpus projects is, first and foremost, the focus MUCH will have on writing as a process. This will be done primarily by including several drafts of a paper, student self-reflective papers, and teacher and peer feedback in the corpus. The student papers included in the corpus range from undergraduate to PhD levels. The corpus also enables analysis of the writing process, rhetorical structures, pedagogical aspects of teaching writing as a process and linguistic structures. Research into the writing process is relevant because it gives, for example, new insights into the role of feedback, drafting and revision in the writing process, which in turn can facilitate new pedagogical developments in the teaching of writing.

The field of writing research, investigating for example peer and instructor feedback is very active, and many studies have been published fairly recently (Beach & Friedrich 2006; Cho, Schunn & Wilson 2006, Cho & Schunn 2007, Hyland & Hyland 2006, Patchan, Charney & Schunn 2009). Many of these studies are, however, studies of L1 writing, often carried out in connection with Writing Across the Curriculum (WAC) and Writing in the Disciplines (WID) programmes in the United States. These studies typically focus on writing in one particular course and therefore often involve few texts, but may contain many peer and instructor comments (Patchan et al. 2009). In this context, one of the contributions of MUCH will be to follow text development and feedback in a greater variety and number of texts. Pedagogical research can therefore make use of MUCH to study the impact of feedback, the development of academic literacy, scaffolding techniques, peer response processes, EFL didactics, etcetera.

Corpus design

The project is still in the very early stages and many decisions are yet to be made on issues such as archiving, tagging, tagging software and interfaces. In this presentation, we give a brief overview of the type of data being collected and give some examples of the issues that may be interesting to study by means of our corpus.

As mentioned above, MUCH will consist of both undergraduate and PhD texts. During 2012-2013, the plan is to collect approximately 400 student papers and 50 PhD texts in three drafts, including peer and teacher feedback. Approximately 150 self-reflective papers will also be included in this version of the corpus. These texts will make up the first version of the corpus, consisting of at least 500,000 words of running text, excluding peer and teacher comments.

In order to show what types of questions can be asked and investigated, a simple example is shown in Figure 1. The box on the left-hand side contains a passage from the first draft of the PhD text, the text inside the circle shows comments from two peer PhD students

on that first draft, and the box on the right-hand side shows the student's revised draft and what kind of changes the comments resulted in (highlighted).

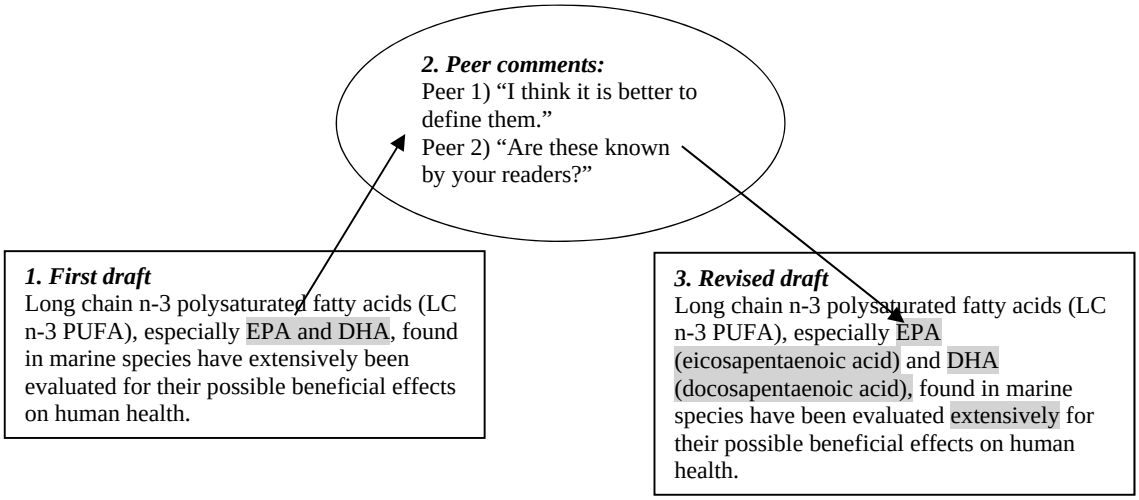

Figure 1. Passage extracted from the first and revised draft of a PhD text in combination with peer comments on the first draft. Parts that differ between the two drafts have been highlighted.

The example shows that the PhD student has decided to change the text by spelling out the full forms of the abbreviations EPA and DHA. The two peer comments concern formal aspects of the texts, and it is possible to tag the comments as such, but it is also possible to make a difference between comments in that one of them is a statement whereas the second one is a question. In this particular case, the two peers have made similar comments, and by tagging comments it is possible to investigate whether students are more likely to change a passage if two students have made similar comments than if only one student has commented on a particular part of the text. Such results would potentially have an influence on the way in which peer work is organized. In figure 1, it is also worth noting that the author has changed the word order by moving 'extensively,' although neither of the peers has commented on this.

Figure 2 displays an example from an undergraduate text where a teacher has commented on the second draft. The comment is at a different level than the example shown in Figure 1, as it concerns the argumentation of the text and the importance of linking to questions and hypotheses.

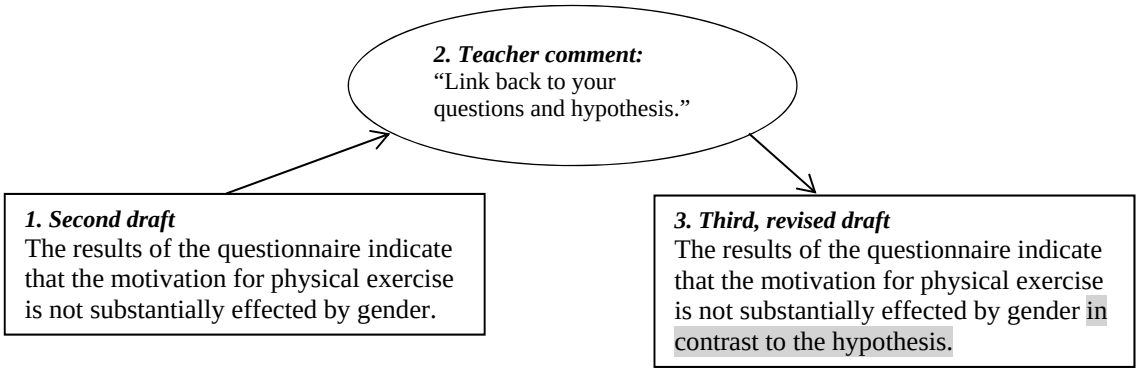

Figure 2. Passage extracted from the second and third draft of a student text in combination with teacher comments on the second draft. Parts that differ between the two drafts have been highlighted.

The change made by the student between draft 2 and 3 is comparatively small, but still illustrates a comment that addresses higher order concerns and that it is possible to tag comments of different types.

Many of the comments from peers as well as teachers are both longer and more complex in terms of what they actually address than the comments shown in Figure 1 and 2. The coding of comments is therefore obviously more complex than the examples suggest. There are also many technical issues that need be resolved concerning the display of results from searches on peer and teacher comments. We are convinced, however, that solving these issues will be worthwhile as it will facilitate many interesting and important investigations.

Already at this point, we foresee that we will need to make decisions on the format for the data in the corpus, as well as to decide on archiving and metadata frameworks. In the work with linguistic and rhetoric annotation of the corpus, tagging systems and categories will need to be inventoried and perhaps partly invented. The latter, we expect, will be the case especially for the tagging of the rhetorical structures we aim to capture in the texts, as the systems for annotating these are less developed than the systems for linguistic tagging. Another challenge is making the writing process available through an interface to the corpus that can visualize the development between the different versions; that is, drafts leading to the final paper and peer/teacher feedback. The general interface through which the corpus will be made public needs to be planned and designed.

The ethical aspects of the publication of the texts, along with informant and other metadata, deserve their own separate analyses and ensuing strategies. Already at this stage, all material gathered through the courses that supply the primary data to the corpus has been volunteered for research and teaching purposes by the students of the courses (the informants). However, further consideration of ethical aspects might be relevant to make the corpus publicly available.

While still in its infancy, MUCH is an attempt at broadening the scope of learner corpora, at the same time as it tries to narrow the gap between writing pedagogy and the use of corpora for teaching and learning purposes.

References:

Beach, R. & Friedrich, T. (2006). 'Response to writing.' In MacArthur, C. A., Graham, S. & Fitzgerald, J. (eds.). Handbook of Writing Research. New York: The Guilford Press.

Berkenkotter, C. & Huckin. T. N. (1995), *Genre Knowledge in disciplinary communication: cognition, culture, power*. Hillsdal, NJ: Lawrence Erlbaum Associates.

Boulton, A. (2009), 'Data-driven learning: reasonable fears and rational reassurance.' *Indian Journal of Applied Linguistics*, 35(1): 81-106.

Boulton, A. (2010), 'Learning outcomes from corpus consultation.' In Moreno Jaén, M., Serrano Valverde, F., & Calzada Pérez, M. (eds.), *Exploring New Paths in Language Pedagogy: Lexis and Corpus-Based Language Teaching*. London: Equinox. 129-144.

Charles, M. (2007), 'Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions.' *Journal of English for Specific Purposes* 6: 289-302.

Cho, K., & Schunn, C. D. (2007). 'Scaffolded writing and rewriting the discipline: A web-based reciprocal peer review system'. *Computers & Education* 48(3): 409-426.

Cho, K., Schunn, C. D. & Wilson, R. W. (2006). 'Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives.' *Journal of Educational Psychology* 98(4): 891-901.

Coxhead, A. (2000), 'A New Academic Word List'. *TESOL Quarterly* 34(2): 213-238.

Flowerdew, L. (2008), 'Corpus linguistics for academic literacies mediated through discussion activities.' In Belcher, D., & Hirvela, A. (eds), *The Oral-Literate Connection: Perspectives on L2, speaking, writing and other media interactions*. Ann Arbor: University of Michigan Press. 268-287.

Flowerdew, L. (2010), 'Using a corpus for writing instruction.' In O'Keeffe, A., & McCarthy, M. (eds), *The Routledge Handbook of Corpus Linguistics*. London, New York: Routledge. 444-457.

Gavioli, Laura. (2005), *Exploring Corpora for ESP learning*. Amsterdam: Benjamins.

Hinkel, E. (2004), *Teaching academic ESL writing: practical techniques in vocabulary and grammar*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hyland, K. (2008), 'Academic clusters: Text patterning in published and postgraduate writing.' *International Journal of Applied Linguistics* 18: 41-62.

Hyland, K., & Hyland, F. (eds), (2006), *Feedback in Second Language Writing: Contexts and Issues*. Cambridge: Cambridge University Press.

Nesi, H, Gardner, S, Forsyth, R, Hindle, D, Wickens, P, Ebeling, S et al. (2005), 'Towards the compilation of a corpus of assessed student writing: An account of work in progress'. In Danielsson, P., & Wagenmakers, M. (eds). *Proceedings from the Corpus Linguistics Conference Series*. Birmingham: University of Birmingham. Date of access: May 29, 2011. Retrieved from: http://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx

Patchan, M., Charney, D. & Schunn, C. D. (2009), 'A validation of students' end comments: Comparing comments by students, a writing instructor, and a content instructor.' *Journal of Writing Research* 1(2): 124-152.