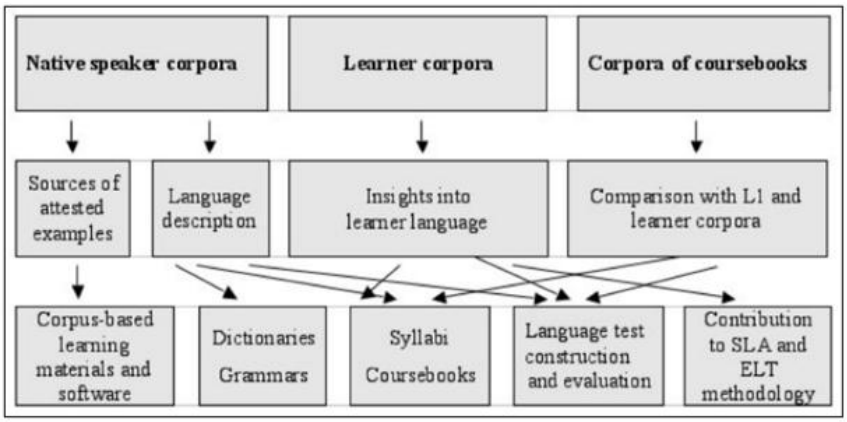
Combining diagnostic and prescriptive corpus functions for ELT:

The HKU CAES learner corpus

Dr. Peter Crosthwaite
Centre for Applied English Studies,
The University of Hong Kong

Introduction

This paper for IELT-Con 2015 looks at the usefulness of native language and learner corpora as a diagnostic and prescriptive tool for English language teaching (ELT). As the use of corpora for ELT may now be described as a 'marriage', rather than a 'fling' (following Gabrielatos, 2005), corpora are now considered as essential tools driving innovation in ELT (Hyland & Wong, 2013). The range of potential uses of both native speaker (L1) and language learner (L2) corpora in ELT are summarised by Gabrielatos (2005) below.



Corpora and ELT (Gabrielatos, 2005)

After describing how corpora are currently being used for ELT purposes, I also describe the construction of a new *HKU CAES learner corpus* that aims to fulfil both diagnostic and prescriptive functions relevant to ELT, particularly for English for academic purposes (EAP).

Use of corpora as a prescriptive tool for ELT

Both native language and learner corpora are now increasingly used in the West and Asia to drive educational research, pedagogy, assessment and publishing. Leech (1997, cited in McEnery & Xiao, 2010) suggests that corpora have both *indirect* and *direct* uses in teaching.

The former – indirect uses - include the creation of dictionaries and reference grammars (e.g. *Longman Grammar of Spoken and Written English* [Biber et al., 1999]), well as the sequencing of course curricula (Thorne, Reinhardt & Golumbek, 2008), and the derivation of course assessment and grading criteria (Barker, 2006, 2010; Hawkins & Buttery, 2009). Increasingly,

the use of corpora for EAP have allowed for greater insights into the type and frequency of linguistic features that are representative of academic writing generally (Wright, 2008; Nesi, 2009; Boulton, Carter-Thomas & Rowley-Jolivet, 2012) or across individual disciplines (Hyland, 2000; Hyland & Tse, 2007), which can be used to enhance the teaching of these disciplines (Reguzzoni, 2013). For example, measuring native-language professional vocabulary alongside wordlists generated from EAP coursebooks can aid the future development of such coursebooks and other course materials (Alexander, 2007; Jones & Durrant, 2010).

The latter - direct uses - typically termed 'data-driven learning' (Johns, 1990) involves teaching students how to use corpora, or employing corpus-based approaches to teaching course content. Such uses should be 'an integral part, rather than an additional option, of the overall language curriculum' (McEnery & Xiao, 2010:24). As an example, corpora are increasingly used to help L2 learners improve their writing, with Quinn (2015) providing an incredibly useful training sequence for L2 writers in the use of concordancing as an alternative to traditional dictionaries, and with Tono, Satake & Miura (2014) training L2 writers to query British National Corpus concordance lines when revising their writing for grammatical and lexical errors, noting improved correction of omission and addition errors. Cotos (2014) compared the use of native language and learner corpora generated from students' own L2 data in the production of linking adverbials over a 10-week period, noting sustained increases in the use of these forms in both groups, and significantly improved post-test scores for the learner group. In each of these studies, despite concerns about time investment and of initial comprehension of concordance data, student reaction to the use of corpora as a prescriptive tool was gauged as very positive.

Corpora as a diagnostic tool – learner corpora

With the exception of studies such as Cotos (2014) above, prescriptive uses of corpora are generally sourced from native language data in general purpose corpora, used as an exemplar for 'appropriate', 'ideal' language use, or at least representative of disciplinary discourses (Hyland, 2000). Regarding ELT, L2 data, which contain numerous grammatical errors and pragmatic infelicity (e.g. Caines & Buttery, 2014; Crosthwaite, 2014), have traditionally been deemed considerably more difficult and labour-intensive to create and analyse for multimillion-word collections. However, Gilquin, Granger and Paquot (2007:328) note: 'What L2 learners need is EAP resource books addressing the specific problems they encounter as non-native writers [...] Yet, hardly any materials writers up to now have taken up the challenge of using learner corpus data.' McEnery & Xiao (2010) suggest this is because 'learner corpora are no longer in their infancy but are going through their nominal teenage years – they are full of promise but not yet fully developed' (2010:19). Milton and Tsang (1991) also note the importance of learner corpora to investigations of L2 interlanguage, in that 'without a reliable index of the degree of difficulty that our students have with the various dimensions of written English such as its lexis, syntax, pragmatics and semantics, we are left to make do with approximations based on impressions, anecdotes and manual counts of small samples' (216).

Using learner corpora generated in a specialised context – in this case, the classroom – researchers and teachers are able to quantify large amounts of L2 data in pursuit of overcoming these challenges, either through comparison with native-speaker data (L2 versus L1) or with different interlanguage varieties (L2 versus L2), a method known as Contrastive Interlanguage Analysis (Granger, 1996). By observing such data in this manner, teachers can 'define areas that need special attention in specific contexts and at different levels of competence, and so devise syllabi and materials' (Gabrielatos, 2005:6), alongside gaining improved insights into the language learning process through the analysis of learning as a product, evidenced in real, authentic L2 production. This is because, as claimed in McEnery & Xiao (2010), 'if learner performance data is shaped and constrained by such a mental process, it at least provides indirect, observable, and empirical evidence for the language acquisition process' (p.18).

A number of useful studies have already attempted to categorise L2 academic interlanguage in this way, including Hyland and Milton (1997), who found that L1 Cantonese speakers have difficulty expressing doubt and certainty in L2 English discourse, and Chuang and Nesi (2006, 2007) who developed a corpus analysing the errors of L1 Mandarin speakers, and then used it to develop teaching materials for academic English students. Crosthwaite (2013, 2014) used the Cambridge Learner Corpus (Nicholls, 2003) alongside a purpose-built hand-annotated corpus of L2 English narrative discourse by L1 Mandarin learners to chart the development of references to person across six L2 proficiency levels. Using a corpus of L2 English argumentative essays from L1 Cantonese speakers, Flowerdew (2006) observed 'signalling' nouns to be problematic for L2 learners, and found a positive correlation between the frequency of signalling nouns and test scores. Finally, a more recent longitudinal corpus study of the type proposed in this paper can be found in Li and Schmitt (2009). They found that one student acquired 166 new lexical phrases over the course of one semester, despite relying on a limited range of phrases overall.

Given the above, it is now increasingly apparent that through corpus analysis, some of the most revealing insights into the efficacy of ELT methodology may be ascertained. In effect, if teachers are able to collect, generate and most of all *quantify* what their learners are actually producing as learner corpus data, the diagnostic advantages of corpus analysis can be used to drive prescriptive goals, while analysis of the results of such prescriptive activity can also provide valuable diagnostic feedback, in a loop of data-driven improvements to product and process.

The HKU-CAES learner corpus – Diagnosing the prescription.

Through careful design, learner corpora can be used to demonstrate - in *quantifiable* terms - the impact of ELT methodology on student performance over time, in effect fulfilling both a diagnostic and prescriptive function if such data is utilized to facilitate data-driven learning for future or even present cohorts of students. However, specialised longitudinal learner corpora are

relatively rare in the corpus linguistics literature, particularly for ELT-related studies. Hence, this section describes the construction of a new learner corpus built specifically to measure the effectiveness of ELT methodology and curricula on actual student production.

Currently under construction, the *HKU-CAES learner corpus* is comprised of data collected from undergraduate students taking their first year of EAP training. It is highly important to investigate these student's initial strengths and weaknesses in English post-secondary education, but, more importantly, to track their development through their early months at university. The pilot study has collected texts from approximately 100 students spanning 300,000 words, which will be extended to 1500 students spanning over 3,000,000 words. Students taking the course currently complete the three following writing tasks:

- A 500- to 600-word diagnostic writing task completed in Week 1. Students are asked to write an (ungraded) essay or report on a single topic.
- An additional 600- to 800-word essay or report submitted in Week 9, worth 15% of the total grade for the course.
- A final written test of 1000-1200 words, again in essay or report form in Week 13, worth 35% of the total grade for the course.

During the semester, the EAP course in question covers units on expressing stance, including counter-arguments and rebuttals, and paraphrasing, units on structuring a complete text, including paragraphing, introductions, conclusions and linking forward and backward between sections, online units on grammar (nominalisation, noun phrases, voice, reference, tenses, hedging, boosting, discourse connectives, prepositions, articles, subject-verb agreement), online units on academic vocabulary, and online units on in-text citations, referencing, and avoiding plagiarism. By tracking student production longitudinally, we are able to determine – in a quantifiable way – whether the provisions of the course outlined above are improved in the learner data in terms of frequency, accuracy (or both).

The corpus is primarily constructed in UAMCorpusTool (O'Donnell, 2008), an incredibly user friendly corpus annotation software that requires little in the way of dedicated scripting knowledge, making it accessible to researchers and teachers alike. One particular advantage of this program lies in its 'autocode' function, allowing for stretches of 'keyword in context' (KWIC) concordance lines to be annotated according to user-specified criteria at the touch of a button. Sketchengine® can also be used on the data for a variety of statistical comparisons, such as a 'word sketch', which lists a word's collocational and colligational distribution across a range of criteria.

Once we are able to diagnose whether our EAP course helped to develop students' production according to the criteria we have proposed, the next step is to use such data in a prescriptive sense. We propose to include, in future course materials, exercises that target the key errors that our students frequently produce. Perhaps more directly, students will have access to the corpus

during in-class and out-of-class writing practice, allowing them to query problematic words and constructions in their own writing and comparing their writing with that of their peers, against 'A' graded exemplars, or against established reference corpora such as the British Academic Written English Corpus. As mentioned, it is hope that this practice would feedback into the corpus itself, allowing for the continuous targeting of student issues with L2 writing into the future.

Conclusion

Corpora have an important diagnostic and prescriptive potential for ELT. By utilising such potential through general purpose and specialised native-language and learner corpora, teachers can have a far more reliable sense of what their students are able to do, and can help shape future development through data-driven improvements to materials and targeted pedagogy.

References

Alexander, O. (2007). Introduction. In O. Alexander (Ed.), *New approaches to materials development for language learning* (pp. 9-14). Oxford, UK: Peter Lang.

Barker, F. (2006). Corpora and language assessment: Trends and prospects. *Research Notes, 26*, 2-4.

Barker, F.(2010). How can corpora be used in language testing? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 633-645). Abingdon, UK: Routledge.

Biber, D., Johansson S., Leech G., Conrad S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*. London: Longman.

Boulton, A., Carter-Thomas, S., & Rowley-Jolivet, E. (Eds.). (2012). *Corpus-informed research and learning in ESP: Issues and applications.* Amsterdam, The Netherlands: John Benjamins.

Caines, A & Buttery, P (2014) The effect of disfluencies and learner errors on the parsing of spoken learner language. *First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, 74–81, Dublin, Ireland, August 23-29 2014.

Chuang, F-Y., & Nesi, H. (2007). GrammarTalk: Developing computer-based materials for Chinese EAP students. In O. Alexander (Ed.), *New approaches to materials development for language learning* (PP. 315-330). Oxford, UK: Peter Lang.

Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL, 26*(2), 202-224.

Crosthwaite, P.R. (2013). An error analysis of L2 English discourse reference through learner corpora analysis. *Linguistic Research,* 30(2), 163-193.

Crosthwaite, P.R. (2014). *Differences Between the Coherence of Mandarin and Korean L2 English Learner Production and English Native Speakers: An empirical study.* Unpublished doctoral dissertation, University of Cambridge.

Flowerdew, J. (2006). Use of signaling nouns in a learner corpus. *International Journal of Corpus Linguistics, 11*(3), 345-362.

Gabrielatos, C. (2005). Corpora and Language Teaching: Just a fling or wedding bells? *TESL-EJ, 8(4),* 1-39.

Gilquin, G, Granger, S & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes.* 6(4), 319-335.

Granger S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Aijmer K., Altenberg B. and Johansson M. (eds) *Languages in Contrast. Text-based Cross-linguistic Studies* [Lund Studies in English 88]. Lund: Lund University Press, 37-51.

Hawkins, J., & Buttery, P. (2009). Using learner language from corpora to profile levels of proficiency: Insights from the English Profile Programme. In L. Taylor & C. Weir (Eds.), *Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment - Proceedings of the ALTE Cambridge Conference, April2008, Studies in Language Testing Series*(31) (pp. 158-175). Cambridge, UK: UCLES/ Cambridge University Press.

Hyland, K. (2000). *Disciplinary Discourses: Social Interactions in Academic Writing*. Edinburgh, UK: Pearson.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing*, 6(2), 183-205.

Hyland, K. & Tse, P. (2007). Is there an 'academic vocabulary'? *TESOL Quarterly*, 41(2), 235-253.
Hyland, K & Wong, L. (2013). *Innovation and Change in Language Education*. London: Routledge.
Johns, T. (1990) From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. CALL Austria, 10: 14–34
Jones, M., & Durrant, P. (2010). What can a corpus tell us about vocabulary teaching materials? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics.* London, UK: Routledge.
Leech, G. (1997) 'Teaching and language corpora: a convergence' in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.) *Teaching and Language Corpora*, pp. 1-23. London: Longman.
Li, J. & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing.* 18, 85-102.
McEnery, A. & Z. Xiao (2010) What corpora can offer in language teaching and learning. In E. Hinkel (ed.) *Handbook of Research in Second Language Teaching and Learning* (Vol. 2). London / New York: Routledge.
Milton, J., & Tsang, E. (1991). A corpus-based study of logical connectors in EFL students' writing: directions for future research. In R. Pemberton & E. Tsang (Eds.) *Studies in Lexis* (pp. 215-246). Hong Kong: The Hong Kong University of Science and Technology.
Nesi, H. (2009). A multidimensional analysis of student writing across levels and disciplines. In M. Edwardes (Ed.), *Taking the Measure of Applied Linguistics: Proceedings of the BAAL Annual Conference* (pp. 81-84). London, UK: BAAL/Scitsiugnil Press.
Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics Conference* (pp. 572–581). Lancaster University: University Centre for Computer Corpus Research on Language.
O'Donnell, M. (2008). Demonstration of the UAM CorpusTool for text and image annotation. *Proceedings of the ACL-08: HLT Demo Session (Companion Volume)*, pages 13–16, Columbus, June 2008.
Quinn, C. (2015). Training L2 writers to reference corpora as a self-correction tool. *ELT Journal, 69*(2), 165-177.
Reguzzoni, M. (2013). Building and using field-specific pedagogic corpora to enhance ESP teaching. In T. Pattison (Ed.), *IATEFL 2012: Glasgow Conference Selections* (pp. 191-193). Canterbury, UK: IATEFL.
Tono, Y, Satake, Y & Miura, A. (2014). The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL, 26*(2), 147-162.
Thorne, S. L., Reinhardt, J., & Golombek, P. (2008). Mediation as objectification in the development of professional discourse: A corpus-informed curricular innovation. In J. P. Lantolf & M. Poehner (Eds.), *Sociocultural theory and the teaching of second languages* (pp. 256-284). London: Equinox.
Wright, A. (2008). A corpus-informed study of specificity in Financial English: The case of ICFE reading. *ResearchNotes, 31*, 16-21.