Corpus of High School Academic Texts (COHAT): data-driven, computer assisted discovery in learning academic English

Róbert Bohát[1], Beata Rödlingová[2], and Nina Horáková[3]

Abstract. Corpus of High School Academic Texts (COHAT), currently of 150,000+ words, aims to make academic language instruction a more data-driven and student-centered discovery learning as a special type of Computer-Assisted Language Learning (CALL), emphasizing students' critical thinking and meta-cognition. Since 2013, high school English as an additional language (EAL) students at the International School of Prague (ISP) have worked with corpora to discover the patterns of English in academic contexts. The positive results of their work with corpora inspired the creation of COHAT, providing a high school level bank of exemplary academic English texts by their native and non-native peers. Our focus is on detecting patterns of correct word choice, syntax and style in student writing.

Keywords: corpus linguistics, discovery learning, learner corpus, critical thinking.

1. Introduction

"Language should be studied in actual, attested, authentic instances of use, not as intuitive, invented, isolated sentences" (Stubbs, 1993, p. 2). In order to help high school students benefit from this approach, COHAT was established by the EAL Department at the International School of Prague.

1. International School of Prague, Praha, Czech Republic; rbohat@isp.cz
2. International School of Prague, Praha, Czech Republic; brodlingova@isp.cz
3. International School of Prague, Praha, Czech Republic ; nhorakova@isp.cz

How to cite this article: Bohát, R., Rödlingová, B., & Horáková, N. (2015). Corpus of High School Academic Texts (COHAT): data-driven, computer assisted discovery in learning academic English. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 71-76). Dublin: Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2015.000312

Since 2013, EAL students at ISP have engaged in a heuristic approach to academic English within the framework of the Applied Linguistics Project (ALP - see our ALP paper in this volume). Students who worked with corpora (InterCorp and BNC) presented their semantic or grammatical discoveries to their classmates, becoming co-teachers in the classroom, which resulted in a lively atmosphere of genuine academic discussion and discovery, confirming the old Latin maxim *Docendo discimus* (By teaching others we learn). The rationale for building a corpus of this type is twofold. First, the existing academic corpora seem to cater predominantly to university level students. Second, most of the high-school learner corpora we found focused on identifying problem areas in non-native speaker texts; the goal of COHAT is to provide high school students with a set of successful academic English texts written by their peers that would focus on detecting patterns of correct word choice, syntax, grammar and style in students' writing. Thus, while recognizing the great value of error detection in learner corpora, we suggest that their value can be enhanced by studying also what the students did well.

2. Method

2.1. Corpus collection and structure

At this first stage, COHAT is relatively small, containing 101 texts with 150,000+ words without annotation, allowing for concordancing, keyword lists, frequency studies, context analysis, and basic collocation studies, using AntConc software (Anthony, 2014). Currently, four discipline-related categories are represented: English and Literature, Social Studies, Maths and Natural Sciences, and Creative Writing (including Speeches & Journalism). Additionally, genres, registers, author's age, gender and mother tongue are recorded in the metadata. The plan is to make it a balanced, representative corpus of student and teacher writing for each high school grade level (i.e. Grade 9, 10, 11, and 12). Following this, grammatical and semantic taggers will be used to annotate the anonymized student texts for further linguistic analysis. As in the university level British Academic Written English (BAWE) Corpus, "only texts that have met departmental requirements for the given level of study" (Alsop & Nesi, 2009, p. 71) were and will continue to be included. The texts are unedited, allowing students to see that some mistakes do not necessarily prevent texts from success.

2.2. COHAT: student discovery in the classroom

How can a learner corpus be used beyond traditional error detection? COHAT and other corpora have been used at ISP also in a constructivist way - as a resource for

detecting the patterns of successful language use across genres, subjects, and registers. Students can use either a corpus, or corpus based data sets prepared by teachers, to analyze and generalize their observations about lexis, grammar, or sentence and paragraph structures. Teachers have used the newly built COHAT to create lesson plans for language discovery activities. A few of these lesson plans have been presented at the English Acquisition and Corpora Building 2015 conference in Pardubice, Czech Republic (more information in Bohát, Rödlingová, & Horáková, 2015, this volume). Below we would like to outline a few preliminary results of COHAT analysis as a starting point for developing a wider set of language discovery lesson plans.

One such area of interest is collocations; here the students are encouraged to use the corpus to get to 'know the words by the company they keep', noticing the nuances of meaning emphasized by habitual co-occurrence of words (Firth, 1957). They also help identify idiomatic and fixed expressions. Figure 1 shows the example of the top two collocates of "largely" that tend to have negative connotations in science discourse. Such studies can inform students' word choice with a view to the prevalent collocations and connotations.

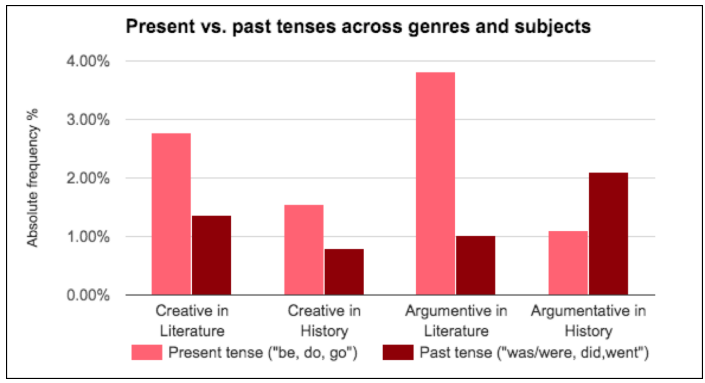
Figure 1. 1R collocates of "largely" (by relative frequency)

1R collocates of "largely" (by RF)

1	**disproportional**
2	**subjective**
3	polish
4	through
5	because
6	termed
7	indicates

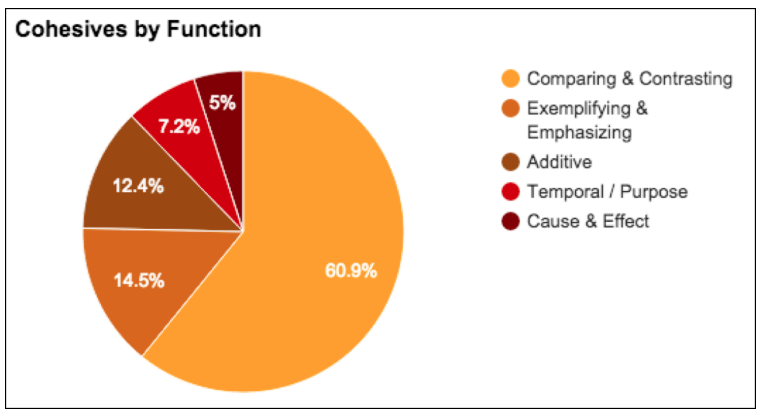
A grammatical aspect of language can be illustrated by the use of present and past verb tenses in writing about history versus fiction. Figure 2 shows a stronger presence of the present tense of selected verbs in literary analysis and of the past tense in history writing. This is an interesting finding because most teachers emphasize the need to use present tense in fiction and past tense only in discussing historical events. Yet, the COHAT analysis shows past tense verbs in successful literary analysis texts. Thus, a corpus turns out to be – inter alia – also an indicator of whether the teachers' explicit preferences are fully and consistently reflected in exemplary student texts. This may be a good stimulus for student and teacher reflection on the rule.

Figure 2. Present vs. past tenses across genres and subjects



Making texts "flow" is often viewed as an abstract, elusive concept. Figure 3 shows a function based analysis of cohesive devices in COHAT with a 60% prevalence of comparing and contrasting cohesives (transitional words or expressions). Students and teachers can reflect: is this a lack of balance in our students' thinking, using comparative analysis at the expense of cause-and-effect relationships or temporal/purpose expressions? Or is it a symptom of an imbalance in the corpus? Could it indicate the popularity of this genre among the teachers, as these are the text types so frequently provided by our colleagues as exemplary? There could be very good (maybe developmental) reasons for the strong presence of comparative and contrastive texts in earlier years of high school education. Either way, having specific data can help teachers and students think critically about their academic language use and inform future lesson plans.

Figure 3. Cohesives by function



3. Discussion

In terms of language pedagogy, corpora allow for a new application of Wittgenstein's (2001) concept of "language games". Applied to natural language acquisition, the game metaphor presents a challenge: language is a game whose rules are not outlined in a user's manual; the native speaker discovers the rules by observation. In language learning contexts, some of the rules are written down in textbooks, but these are often difficult to understand and internalize. Worse still, thinking about memorized rules of grammar before uttering a sentence typically slows down student communication.

On the other hand, computerized corpora can serve as a playground of sorts for students to observe large quantities of language in use and at least partially make up for the lack of years of constant exposure available to native users. Just as an observer of a chess game played by others can soon start making observations and generalizations about the use of individual pawns, a language learner can observe texts and contexts in a corpus and literally play with the language, similar to what the great 17th-century Czech educator Jan Amos Comenius proposed as Schola ludus – school (learning) by play. This approach resembles natural language acquisition in making learning emerge from meaningful, playful activities.

4. Conclusions

A corpus of exemplary high-school texts going beyond error detection seems to be missing or not readily available. COHAT is designed to serve that purpose and enable students to conduct their own research into the language of academic writing, and with teacher support rediscover 'the rules of the language game' for themselves. Its texts represent an achievable goal for high school students, and the pedagogical approach proposed here encourages inferential, data-driven discovery learning. Additionally, it has proven to be a useful indicator of (dis)harmony between teaching theory and student practice.

A profound definition of learning says that it is "constructing knowledge in collaboration with others", a guided rediscovery of knowledge (Wells, 2001, p. 176). The use of corpora in Academic English classrooms has shown that this rediscovery can be applied to gaining linguistic knowledge, too. It is true that "language looks rather different when you look at a lot of it at once" (Sinclair, 1991, p. 100). High school students of any academic language can benefit from being able to 'look at a lot of language at once' through the lens of a corpus, distinguishing language use by genre, subject or register, for "he who distinguishes

well, teaches well", argued Comenius (1948, p. 162). Even better, corpora can be used to help students "distinguish well" through a set of computer assisted game-like learning activities, increasing their engagement by the element of discovery and adventure.

5. Acknowledgements

We would like to thank Lawrence Hrubes (ISP EAL department chair), ISP Administration, as well as doc. Mgr. Václav Cvrček, Ph.D., Mgr. Michal Křen, Ph.D., and Mgr. Lucie Chlumská from the Institute of the Czech National Corpus for their invaluable professional and logistical help. Róbert Bohát would like to thank his wife and family for their support without which most of this work could not have been completed.

References

Alsop, S., & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora, 4*(1), 71-83. doi:10.3366/E1749503209000227

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/

Bohát, R., Rödlingová, B., & Horáková, N. (2015). Applied linguistics project: student-led computer assisted research in high school EAL / EAP. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 65-70). Dublin: Research-publishing. net. doi:10.14705/rpnet.2015.000311

Comenius, J. A. (1948). *Vybrané spisy Jana Amose Komenského. Svazek 1. Praha*. Státní pedagogické nakladatelství.

Firth, J. R. (1957). *Papers in linguistics 1934-51*. Oxford: Oxford University Press.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stubbs, M. (1993). British traditions in text analysis – from Firth to Sinclair. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: in honour of John Sinclair* (pp. 1-33). Amsterdam: John Benjamins Publishing Co. doi:10.1075/z.64.02stu

Wells, G. (2001). Action, talk, and text: learning and teaching through inquiry. New York: Teachers College Press.

Wittgenstein, L. (2001). Philosophical investigations (3rd ed.). Oxford: Blackwell Publishing Ltd.

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; info@research-publishing.net

Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy
Edited by Francesca Helm, Linda Bradley, Marta Guarda, and Sylvie Thouësny

Typeset by Research-publishing.net
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-28-5 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-29-2 (Ebook, PDF, colour)
ISBN13: 978-1-908416-30-8 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2015.