

ENGLISH FOR ACADEMIC PURPOSES

REFLECTIONS, DESCRIPTION & PEDAGOGY

SIMONE SARMENTO, ROZANE REBECHI,
MARINE LAÍSA MATTE (ORG.)

e for learning English.</s></s>This may include EAP (on, Canada so that a student can complete our EAP (/ students.</s></s>I'm TESOL certified to teach EAP (ents is:</s></s>This series from award-winning EAP (, not apply; however, six credits of college-level EAP (or at Emory University's Candler School of Theology (y that ... Continue reading →</s></s>OXFORD EAP (rted my second year of teaching at BU with the EAP (hing English on the BU campus in the EAP program (is article provides a guide to the award-winning EAP (Edward de Chazal explains the challenges that EAP (ses, and adjunct professor for E.</s></s>A.</s></s>P (interests include second language acquisition, EAP (l is required.</s></s>Students take prerequisite EAP (onventions.</s></s>Despite the efforts of many EAP (ernational students at colleges and universities EAP (survey.</s></s>Theoretical Background</s></s>EAP (

), which prepares students at tertiary level for further a) Program (Level 10 with 80%) and then enter the univ) which means that I must be knowledgeable in all ac) author Aylin Graves provides a set of lesson plans to) coursework taken at Florida SouthWestern State Col program).</s></s>He is also academic director of UGA) B1+ INTERMEDIATE - components</s></s>This diss) program.</s></s>I teach level 2 writing every morning)</s></s>Classes consist of International students for) series from author, Aylin Graves.</s></s>Approaches) learners face, and what teaching staff and lecturers r) courses.</s></s>She has spent many hours in the cla) , translation, interpreting, education quality assessme) courses in reading, listening, writing, and research br) researchers and practitioners to provide support for r)) ELT (Enhanced Language Training) ESP (English for) researchers, such as Christison and Krahnke, 1986;

ENGLISH FOR ACADEMIC PURPOSES

REFLECTIONS, DESCRIPTION & PEDAGOGY

**SIMONE SARMENTO
ROZANE REBECHI
MARINE LAÍSA MATTE
(ORG.)**

Porto Alegre • 2024 • 1ª edição

editora
**ZO
UK**

Conselho Editorial

Cristiane Tavares – Instituto Vera Cruz/SP
Daniela Mussi – UFRJ
Idalice Ribeiro Silva Lima – UFTM
Joanna Burigo – Emancipa Mulher
Leonardo Antunes – UFRGS
Lucia Tennina – UBA
Luis Augusto Campos – UERJ
Luis Felipe Miguel – UnB
Maria Amelia Bulhões – UFRGS
Regina Dalcastagnè – UnB
Regina Zilberman – UFRGS
Renato Ortiz – Unicamp
Ricardo Timm de Souza – PUCRS
Rodrigo Saballa de Carvalho – UFRGS
Rosana Pinheiro Machado – University College Dublin
Susana Rangel – UFRGS
Winnie Bueno – Winnieteca

copyright © 2024 Simone Sarmento, Rozane Rebechi, Marine Laísa Matte

Projeto gráfico e edição: Editora Zouk

Revisão: Simone Sarmento, Rozane Rebechi, Marine Laísa Matte

Imagem da capa: SKELL

Dados Internacionais de Catalogação na
Publicação (CIP) de acordo com ISBD

Elaborado por Wagner Rodolfo da Silva - CRB-8/9410

E58

English for Academic purposes [recurso eletrônico] : reflections,
description e pedagogy / organizado por Simone Sarmento, Rozane Rebechi,
Marine Laisa Matte. - Porto Alegre, RS : Zouk, 2024.

268 p. ; ePUB.

Inclui bibliografia.

ISBN: 978-65-5778-135-7 (Ebook)

1. Linguística. I. Sarmento, Simone. II. Rebechi, Rozane. III. Matte, Marine
Laisa. IV. Título.

2024-175

CDD 410

CDU 81'1



direitos reservados à

Editora Zouk

r. Cristóvão Colombo, 1343 sl. 203

90560-004 – Floresta – Porto Alegre – RS – Brasil

f. 51. 3024.7554

www.editorazouk.com.br

Creating a local learner corpus: Insights on project design and data analysis from the pilot phase

Sandra Zappa-Hollman (UBC-CA)

Alfredo Afonso Ferreira (UBC-CA)

Greta Perris (UBC-CA)

Simone Sarmiento (UFRGS)

Marine Laísa Matte (UFRGS)

Laura Baumvol (UBC-CA)

Introduction

A learner corpus (LC) is a principled collection of texts produced by additional language learners. These texts are collected and systematically organized electronically to allow for a range of teaching and research applications. Learner corpora have been typically created by academics or publishers for so-called “delayed pedagogical use” (i.e., not necessarily for the immediate benefits of those students sharing their writing samples), as well as for research purposes; that is, for contributing to theorization in additional language acquisition and applied linguistics through identification of patterns in learner language. More recently, however, a growing number of LCs have been created locally by researcher-practitioners for “immediate pedagogical use” in their specific institutional contexts (Granger, 2009, 2015), leading to data-driven enhancements in curriculum development, teaching, and learning.

The LC project we report on here was designed to systematically collect and access large samples of our students’ writing for relatively immediate pedagogical application. Over time, this resource is meant to better track writing development within and across student cohorts and identify patterns of variation at larger scales such as across disciplines, language background of learners, and instructional programs. This scope of interest

across teaching and research is indicative of the close relationship between them in data-based learning. In addition to helping us systematize access to student texts for research purposes, our LC is also designed to inform curriculum development and instructional practice.

When we embarked on this project, our team represented a range of expertise and background knowledge that enabled us to envision the overall objectives and structure of our LC. Yet it was evident early on that creating a successful local LC would require effort and a steep learning curve. This chapter reports on some of the key choices we made as we designed and implemented the pilot phase of the LC project. Some challenges we had to overcome and important considerations we made in relation to technological and logistical aspects. And to illustrate the potential benefits to teaching and research in our context of even a small dataset from the pilot phase of the project, we also present the results of an analysis of comparative discourse in student expository writing. We close the chapter with reflections synthesizing what we have learned from the pilot phase and outline on our following steps.

The VanCor Project

The Vantage College Corpus of Student Texts Across Disciplines (henceforth, VanCor) project that to create a systematic and searchable online repository of student written assignments. VanCor is conceived as a resource for faculty at Vantage College (VC) in The University of British Columbia (UBC) have easy access to written assignments that students engage in across a range of disciplines in first year programs. VanCor has the potential to be relevant for research, data-driven curriculum development, instructional materials development, and program evaluation purposes.

Institutional Context

Launched in 2014, VC is a unit at UBC that offers first year programming for international English as an additional language (EAL) speakers whose proficiency is slightly below the university's English language admission standards for direct entry. At the time of data collection, three program options were available: first-year Bachelor of Arts, first-year Bachelor of Engineering, and first-year Bachelor of Science. Program faculty include a team of English For Academic Purposes (EAP) instructors who work with disciplinary faculty seconded to VC from their respective departments in Arts, Engineering, and Science.

VC offers instructional programming tailored to support of students' transition into the second-year of their bachelor's degree at UBC. VC programs are characterized by a cohort-based model and standard timetables, providing a coordinated curriculum that includes content-focused and language-focused¹ credit-bearing courses. Thus, alongside their program-specific courses, students receive general EAP and discipline-specific English instruction. After successfully finishing their first year at VC, students continue as second-year students in their respective faculties. The program expands the usual two academic terms of first year to three academic terms, totaling 11 months of instruction. This time extension accommodates the required disciplinary courses in the respective programs of study as well as VC-specific programming aimed at scaffolding students' linguistic, cognitive, and skills development as apprentice multilingual scholars.

The custom-designed programming includes an introductory research methods course with an application component that engages students in a small group research project they eventually present at an annual

1 VC uses an integrated language and content approach which views the learning of language and subject area knowledge as inseparable and mutually constitutive. We use "content-focused" and "language-focused" as shorthand to refer to what otherwise are also referred to in the literature as "subject, or disciplinary" courses versus "language" courses. Yet we view both types of courses as involving both content as well as language. To try and foreground this relationship between language and content, we classify these as courses that place an emphasis or focus in either of the two, based on what most course learning outcomes stipulate.

student-led capstone conference. To explicitly support their academic (general as well as discipline-specific) language and literacy, students complete academic English courses informed by Systemic Functional Linguistics as well as have access to on-demand academic English support via writing consultations.²

These multiple, relatively uncommon, aspects of the programs at VC make it an attractive context for researching learners' language characteristics, use, and development. In what follows, we recount the genesis of the international collaboration that led to the VanCor project.

International Collaboration: A Brief History

The VanCor project brings together researchers and educators from UBC in Canada, and the Federal University of Rio Grande do Sul (UFRGS) in Brazil. The genesis of this project was in late 2019, over conversations amongst Simone, Alfredo, Laura, and Sandra, about ways to collaborate around a project of mutual interest. Since one of the mandates of VC is to serve as a living lab for pedagogical and research innovation, designing a research project with the goal of supporting activities such as curriculum development and design of student tasks seemed most fit and appealing. Given Simone's expertise in LC development and the desire from VC members to create an institutional learner corpus, our group decided to embark on a project, seeing the potential benefits of the international collaboration. By early 2020, we had obtained competitive funding via a Social Sciences and Humanities Research Council (SSHRC) institutional grant. This funding supported the hiring of our two graduate research assistants from UBC. In what follows, we provide an overview of the project sequence and key stages.

2 For further details on the Vantage program, see Zappa-Hollman & Fox (2021), Ferreira & Zappa-Hollman (2019), and Zappa-Hollman (2018), as well as the Vantage College website: <https://vantagecollege.ubc.ca/program-overview>

Project Timeline

The pilot phase involved four stages (Fig. 1).

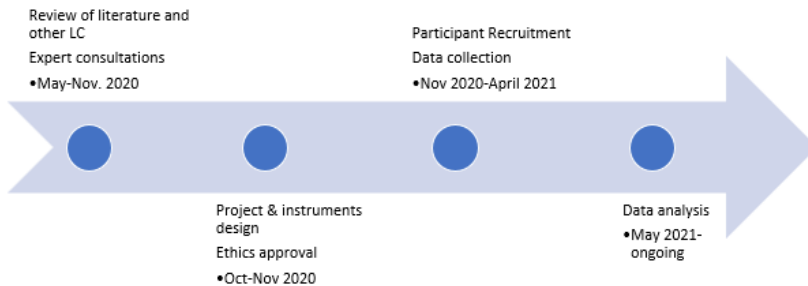


Figure 1. VanCor project Timeline

The first stage involved an extensive, updated review of the literature on learner corpora, with a focus on the creation and uses of LC for research and pedagogical applications. This literature review was complemented with consultations with experts in learner corpora in university contexts outside of Canada as well as with consultations with UBC librarians with expertise in data management.

The second stage involved defining the scope, objectives, procedures, timeline, and developing the data collection instruments. To collect the learner texts that our project participants were willing to share with us, we used a survey (hosted on the Qualtrics survey tool). This survey, reproduced in Appendix A, included a section for collecting demographic data about the participants, a second section about assignments information and for uploading assignments (up to 15 files), and a third section inviting participants for a debriefing interview. The interviews aimed to gather feedback from the students about their experience participating in the study (i.e., completing the survey), potentially offering deeper insights about the process of writing their assignments. To complete the survey, participants had to first provide their informed consent via a form included at the start of the survey. The survey also included the request for student

consent to the collection and use of their data. At this stage we also applied to the institutional ethics board for approval to conduct this pilot study.

The third stage involved participant recruitment and data collection. This stage spanned six months and took place virtually³ in two courses taught by two instructors who are also members of this project team. In late November 2020 (end of our Fall term), we recruited participants in one section⁴ of an academic writing course taught by Laura Baumvol in the Arts program and collected texts from this class until January, 2021. At the start of the Winter term, we recruited participants from two sections of an adjunct course taught by Alfredo Ferreira that links EAP instruction to courses in the Science program. The recruitment was carried out by the two graduate research assistants during a 15-minute class visit of a synchronous session when the instructors were not present. During this visit, the students were introduced to the project through a 5-minute video with an overview of the project goals and a description of what participating involved. This was followed by some Q&A time in case prospective participants had any queries. After the class visit, a link and QR code to the survey was posted as an announcement on the course learning management system sites.

In total, we collected nine assignments and two sets of instructions, and conducted two interviews were conducted; these took place once the final grades for their respective classes had already been awarded. Following data collection, the fifth stage involved data preparation and data analysis. To protect the identity of participants and systematize the process of data management, we assigned unique identifiers to each text and instructions, and removed all personal identifying information prior to starting with data analysis. Next, we used a metadata coding sheet to describe the relevant context and genre of each text. We developed our text metadata coding sheet partly based on a similar resource from Graves and Hyland (2017) with some adaptations for our context and project purposes. The

3 Since our project was carried out during the Covid-19 pandemic, all research activities – including recruitment and data collection – were carried out online.

4 Each course section has a student registration of 25, maximum.

coding sheet can be found in Appendix B⁵. For this classification, we are drawing on Systemic Functional Linguistic theory. Section 5 includes an illustration of the analysis of corpus data for use in research and instruction.

Key Reference Literature

As mentioned above, we consulted canonical texts on learner corpora (Granger, 2002, 2009, 2015; Gardner & Nesi, 2013; Römer & O'Donnell, 2011) to gain insights on types of data to collect, steps, and sequencing to follow, as well as tips to avoid common pitfalls and minimize challenges in data retrieval and analysis. Recent articles focusing on the process of designing and implementing a LC were helpful to learn from insights the authors gained through trial and error.

For instance, Granger et al's (2020) *International Corpus of Learner English* (ICLE)⁶, which is composed of texts written by upper intermediate and advanced learners of 25 different language backgrounds offers an excellent model for gathering metadata on the texts that allow for an in-depth view of both the learners and the tasks.

Some projects have expanded their scope to provide additional types of resources to assist with writing research, support instructors' professional development, and train those intending to design and use an LC. Two such corpora we found impressive in this regard are the *Multilingual Academic Corpus of Assignments: Writing & Speech* (MACAWS) and the *Corpus & Repository of Writing* (CROW), both with Dr. Shelley Staples as a lead investigator. MACAWS (Staples et al., 2019) is an ongoing building corpus of assignments written by students enrolled in language programs at the University of Arizona. CROW (Staples & Dilger, 2018) contains texts that L1 and L2 first-year undergraduate students write in their composition classes in three universities in the US. Access to these resources is available by requesting registration to their customized websites. Once registered

5 This genre classification system will be revised as we collect more texts from different genres.

6 <https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html>

with the MACAWS website, for example, we were able to access a repository of pedagogical materials associated with the assignments, such as syllabi, assignment sheets, lesson plans, and instructional materials; and language learning activities in Portuguese and Russian designed based on the language patterns that emerge from the corpus. In CROW, we also accessed demographic data and a repository of resources intended to help others – like us – with the design and use of LCs. The resources shared in these two projects guided our decisions about several aspects of our own project. For example, the demographic information helped us further refine the kind of metadata we would collect. The corpus helped us reflect on whether collecting drafts of our students’ final writing was useful for our project. The pedagogical resources provided several suggestions for ways in which to involve other VC instructors in the next phases of our project.

Another corpus that informed our pilot is the *The Civil Engineering Writing Project* Conrad (2017)⁷, led by Susan Conrad at Portland State University. The corpus includes student and practitioner writing in the field of engineering, as well as an impressive collection of open-source instructional materials for use by course instructors and students in self-study. The studies that emerged from the project (reviewed below) relate to genre and linguistic analysis, grammar and mechanics errors, as well as holistic evaluations of writing effectiveness. The genre classification in this corpus helped us reflect on the way we would approach genres in our own texts.

We also drew from “Writing assignments across five academic programs” by Graves (2017), a chapter in an edited book by Canadian researchers who created a corpus of undergraduate student assignments. From this resource we used the writing assignments coding guide, which served as the basis for our coding guide. This coding sheet is used to record standard information (e.g. genre, length of text, topic, grade, etc.) collected about each text, which is then entered into the web-based application that compiles them and creates reports. In addition to guiding our coding procedures, the process of adapting the coding sheet became an opportunity for our team to revisit and adjust as needed the goals and scope of the project.

7 <http://www.cewriting.org/>

Alongside canvassing websites and scholarly publications, we also reached out to a number of the scholars who led those works, primarily with questions related to preferred communication and collaboration practices within their research team, and questions related to data collection and management.⁸ The guidance that was generously offered extended well beyond these questions. The scholars candidly shared their experiences of learner corpora development (e.g., MACAWS, CROW) and lessons learned along the way. They highlighted important yet often overlooked aspects of corpus development such as steps to ensure a sufficient number of texts are collected and advised starting small, staying focused on the scope, which may involve starting with a smaller project before scaling it up. Based on these insights, we adjusted our timeline for collecting the texts, and decided that in the scaled-up version of our project we will not provide monetary incentives for participation (as these can prove challenging for distribution as well as add significant cost to the project). These projects also provided access to a wealth of resources encompassing the lifecycle of a corpus-building project, from detailed information on developing the backend of the corpus, such as the database structure, automated tools, indexing, text-processing tools, and illustrations of how corpora can be used to create relevant pedagogical materials.

Drawing on the LC community of practice helped us reflect on our research questions and practices with experts in the field, make informed choices that strengthened our project, and enabled us to refine the project and move forward with heightened confidence. We also consulted experts in digital scholarship through workshops that provided crucial training on the choice of digital tools available for project management, data collection and storage, and the dissemination of project outcomes. These training sessions also introduced us to institutional norms and best practices pertaining to handling sensitive research data (e.g., institutional requirement to store data on Canadian servers).

⁸ We are extremely grateful for the generosity of our colleagues from the Corpus Linguistics field who have kindly shared their knowledge with our team.

Analysis of selected data from pilot project

Insights into student writing based on quantitative analysis of a large sample is a key goal of local learner corpora collected for immediate pedagogical purposes. While a corpus that is sufficiently large for quantitative analysis would have been a welcome outcome of our pilot project, as mentioned earlier, the first aim of the project was to test approaches to data collection in our context. Having described the steps we followed in designing, collecting, and storing the data for our project, we now focus on a small subset of four texts collected from Science students to illustrate how even such a small sample can inform teaching, research, and the VanCor project in valuable ways.

The contributions to instruction of a very limited sample of student texts from the same instructional setting can be likened to those of qualitative, case study analysis (Duff, 2008), the primary overlap being that the data emerge from a specific, well-defined context. As with case study data, such pilot corpus data suggest hypotheses about learner practices that can be subsequently explored in a wider study, targeted for collection in larger corpora, and, on the teaching side, can inform the development of instructional materials to test in classrooms for fit with student needs and interests.

Given the interest in forming hypotheses and developing instructional materials from the pilot data, two aspects of the data come directly into play. First, it is important to recognize that the texts are not representative of the Science program cohort or all students in the class: these are relatively successful texts voluntarily submitted by four high-performing students within the top 10% of the class. This quality of these data point to a weakness in the opt-in approach to the collection of student texts: generally, high-performing students submitted writing assignments that were also high-performing in terms of the grade received.

The second aspect of the data that inform their use for instruction is the nature of the writing undertaken in this assignment and our focus on pedagogical application. The students wrote comparative discussions, approximately 1,400 words in length, across three drafts with instructor and peer feedback. The student-writer selects two scientific theories, concepts,

or approaches in the history of science to compare in relation to a specified criterion. This critically-engaged discussion typically concludes with claims about the different motivations for these concepts in the history of science⁹.

Correspondingly, instruction focused on expository genres, specifically comparative discussions in the history of science. Comparison is a semantic domain relevant in the discussion assignment as well as the first-year Physics, Chemistry, and Mathematics textbooks used by the students, as it is in science discourse more widely. The instructor in this case, Alfredo, observed that students were frequently challenged when using comparative language in their reports, such as from Chemistry labs, as well as in longer writing assignments. This particular discourse analytic research out of VanCor arose from an interest in developing materials that would help address this observed need for instructional materials in the history of science module of the VC science Content and Language Integrated Learning (CLIL) course. Students in the Vantage science stream received no other explicit instruction in the language and functions of comparison.

Qualitative analysis of these texts following SFL theory (Halliday & Matthiessen, 2014) led to a number of instruction and research-worthy insights into the functions of comparison in historical expositions and academic writing more generally. Table 1 outlines the functional range of comparative language identified in student writing across metafunctions (i.e., organization, interpersonal positioning, and representation) and some subfunctions. For background on the one more technical subfunction listed in the table, theme, understood in SFL as the informational point of departure for the clause, see Kang (2016). Within the function of representation, the genre-specific distinction between focal and non-focal compared things is explained below.

⁹ For a relevant outline of the development of disciplinary literacy practices in history, see Coffin (1997).

Meta-functions	Sub-function	Example of Comparative Language in Students' History of Science Writing
Organization	Title	A Comparative Exposition of Celestial Mechanics and Quantum Electrodynamics in relation to the Description of the State of Motion. (Text 09 - Science)
	Thesis statement; topic sentence	In the exposition that follows, phlogiston theory and oxygen theory are compared from macro and micro perspectives in studying science. (Text 10 - Science)
	Theme	Neo-Darwinism places greater emphasis on natural selection, whereas eugenics affirms that artificial selection is required to conserve the useful features of individuals (Paul, 2013). [...] These contrasts will be further discussed within the section below. (Text 07 - Science)
Interpersonal Positioning	Hedge: claim	A more detailed exploration of the kinematic relation between two or more objects in macro and micro perspectives is provided to consider the difference between the types of acting force. (Text 09 - Science)
	Hedge: disciplinary category	"better adapted individuals" can be described as a group of organisms with higher reproductivity which enables their "more useful" genetic characteristics to pass onto their offspring and onto the future generations, whereas "less adapted individuals" are less likely to survive (Abbey & Abalaka, 2011). (Text 07 - Science)
Ideational	Comparison of focal things	Comparing the symmetrical aspect of nature has the possibilities to predict the existence of unknown materials or phenomena in the universe (Capra, 1975). (Text 08 - Science)
	Comparison of non-focal things	A more detailed exploration of the kinematic relation between two or more objects in macro and micro perspectives is provided to consider the difference between the types of acting force. (Text 09 - Science) Both observation and experiments are indispensable in studying science, making science more rigorous and accurate (Ainsworth et al., 1991). (Text 10 - Science)

Table 1. The functional scope of comparative language in high-performing expositions in History of Science by first-year science students.

These data highlight several important features of comparative language that can help develop hypotheses about this area of discourse for use in teaching and research. The main finding is that comparative language

realizes all three main metafunctions and various subfunctions. An interesting example of this is within the function of representation (technically in SFL, the experiential function), which indicates two levels of focus when analyzing genres that explicitly set out to compare things: the comparison of two or more things in focus in the comparative text, and the comparison of everything else for purposes such as organizing ideas for the readers, that is the comparison of non-focal things. The latter function arises, for example, in comparing relative degrees of information detail across the text, where the writer signposts “a more detailed exploration of... is provided”. This finding indicates that comparison is both a defining feature of some genres and a more broadly functional resource in academic discourse.

These corpus data also help qualify the comparative exposition as a useful genre for understanding the development of student writing. This claim is based on the wide functional scope of comparative language, its Field (realized in the lexicogrammatical choices of for representing ideas, Tenor (interpersonal positioning), and Mode (textual organization) (on these register variables, see Halliday & Matthiessen, 2014). Such a map helps us to chart trajectories of development of language and academic writing within and across functions by focusing on comparative language. In this way, the data also lend validity to the assignment in relation to the course learning objectives which aspire for development across the three metafunctions.

In relation to language and writing development, it is worth noting the potential of extending this map. This opportunity has arisen in the transcript of a visiting lecture (which led the history of science module in the course) by an established historian of science¹⁰. The lecture includes instances of comparative language used for engagement (superlative/hyperbole used to bait readers into the counter-argument before arguing against it) and politeness through reverse polarity (reference to an unreliable academic source as “not the most impartial judge”). The extension of the semantic potential of comparative language suggested by the practices

10 The transcript referred to here comes from a lecture on Ancient Greek protoscience delivered by Dr. Sylvia Berryman, Philosophy Professor at UBC.

of a more mature scholar shows how comparative language can realize increasingly fine-grained functions in accordance with disciplinary and linguistic development, illustrating Halliday's (1993) conception of language development as increasing one's registerial repertoire or capacity to mean across situated contexts; for discussion in advanced language development, see Matthiessen (2006). These insights indicate potential directions for researching language and writing development in this context.

Moving to a lexicogrammatical view of comparison in History of Science arguments, an analysis of the comparative lexis from the history texts in the pilot corpus yielded the results shown in Table 2 below. The word lists on the right-side columns are classified by grammatical and semantic/functional units, subunits, and whether the words instantiate the semantic domains of similarity or difference. The ordering of grammatical units from nominal group (noun phrase in traditional grammar) at the top of the table down to verb, adverbial, and conjunction at the bottom is motivated by the degrees of information density afforded by these units (i.e., from most abstract and/or general to most concrete) per the concept of ideational grammatical metaphor (Ferreira, 2020; Halliday, 1998). The wordlist in each of the subunit categories are ordered from most to least frequently occurring with the number of tokens listed on the right-hand column.

Grammatical/ Functional Unit		Subunit Similarity	Instances in Pilot Corpus		#
			Difference		
1	Nominal Group / Participant	Head Noun / Thing	comparison/s		8
			similarities		8
				difference/s	7
				contrast/s	2
			alignment		1
				opposite	1
			superiority	1	
		Adjective & premodifier/ post-pointer, describer		different	8
				opposite	7
				better	3
				greater	3
			both	better adapted	2
			comparative	broader	2
				higher	2
				more likely	2
			similar	deeper	1
			corresponding	less adapted	1
				less likely	1
				more appealing	1
			same	more like	1
				more predictable	1
			more regular	1	
			more useful	1	
	opposing	1			
	proportional	1			
	superior	1			
2	Verb / Process	Relational process	overlap		1
			share		1
		Material & other process	compare/d		9
			comparing	contrasts	3
				distinguished from	2
	correlated with		2		
3	Adverbial / Circum- stance		also	by contrast	1
			like	in contrast with	1
			similarly	more frequently	1
4	Conjunc- tion/ Relator		as	while	4
			as	whereas	2
				however	1
				rather than	1

Table 2. Comparative lexis by grammar and function in four high-performing History of Science expositions in 1st-year EAP

As can be seen, the tokens of comparative language cluster significantly in the nominal group (e.g. “These **contrasts**”; “A **more detailed** exploration of the kinematic relation between two or more objects in macro and micro perspectives”). This result can be understood to reflect the relatively high functional load of the nominal group in academic writing especially with regards to the specification of concepts and foci that is associated with disciplinary writing development in university (Duff et al., 2015). Unsurprisingly, abstract concepts involving comparison are central to texts and genres that set out to compare historical theories in science.

The finding of high frequency of comparative language in nominal groups relative to its use in more dynamic processes (verbs), circumstances (adverbs) and logical reasoning (conjunctions) points to a need for additional attention to this role of comparative language in construing abstract concepts in writing instruction. A cursory examination of two popular EAP writing textbooks, both fourth editions (Oshima & Hogue, 2005; Blanchard & Root, 2017), highlights a potential emphasis on the latter dynamic, syntactically more complex and material meanings, while the more frequent realizations of abstract concepts involving nominal groups receive little explicit attention. The “comparison signal words” recommended as useful for comparative writing in one of these textbooks, shown in Table 3, illustrates this tendency:

Comparison Signal Words
Transition Words and Phrases: similarly; likewise; also; too
Subordinators: as; just as
Coordinators: and; both... and; not only... but also; neither... nor
Others: like (+noun); just like (+noun); similar to (+noun); (be) like; (be) similar (to); (be) the same as; (be) the same (be) alike; (be) similar; to compare (to/with)

Table 3. Words and phrases used in comparisons recommended by popular EAP writing textbook (Oshima & Hogue, 2005: 116-117)

According to this edition of the textbook, students should focus their attention on realizations of comparison for these functions of logical ordering and transition with minimal attention given to elements of the nominal

group (noting that the “+noun” elements under “Others” do not themselves realize a comparative meaning). Such an emphasis does not align with the functional distribution of comparative language in the sub-corpus of high-performing texts in the History of Science.

These results suggest potentially useful insights for research and instruction. We have found that the semantic scope of comparison encompasses a wide functional range of language: ideational, interpersonal, and organizational meanings, and various sub-functions of these such as evaluation, affect and multiple scales of text organization including signposting through topic sentences and various cohesive devices.

Given the wide functional scope and grammatical realizations of comparative language in the comparative exposition genre, a relatively holistic perspective on language and writing development in EAP contexts can be operationalized by focusing on comparative language in this genre. The same results suggest various corpus-based approaches and tasks for instructional curricula involving comparative and related genres of academic writing. In these and other ways, the focus on a few student texts within a relatively specific written genre has yielded useful insights to apply to teaching, research, and the next phases of the VanCor project.

Current status and next steps

Through our collaboration on the VanCor project, the team has taken the first steps in designing, compiling, storing, and applying learner corpora: reviewing the literature, consulting with experts, piloting the various sub-tasks involved in data collection, and analyzing the results. These experiences in the pilot phase of the project will inform the next phase of the project.

Our efforts to disseminate our ideas and experiences range from the local to the global. We introduced our LC project and preliminary findings to our VC program colleagues with the aim of generating interest in collaborating on the larger scale of the project through realizing its potential for curriculum development, instruction, and research. Additionally, we have engaged in dissemination efforts, which include presentations at

professional organization annual conferences¹¹, with the intent of sharing the insights gained from our pilot and sharing our preliminary findings.

As for the next steps in VanCor itself, we plan to implement the project by inviting all VC instructors as collaborators and thus expand the nature of the student texts included in the LC. A higher number of instructor collaborators across all VC programs will allow us to collect texts from, ideally, all courses included in first year programs. This scope of text types will result in a diversity of genres across several disciplinary fields, expanding the potential contributions of the corpus to research.

Acknowledgments

The UBC graduate research assistant positions for this project, fulfilled by Greta Perris and Sara van Dan Acker, were supported by an institutionally administered Social Sciences and Humanities Research Council (SSHRC) grant (UBC Explore SSHRC grant), for which we are very grateful. Simone Sarmento would like to thank the support of CNPq Productivity Grant and CAPES Print. Our gratitude also goes to the students who participated in this pilot phase of our project, generously sharing samples of their writing. We would also like to thank Brian Wilson, Curriculum Manager at VC, for his feedback and advice on survey design; Dr. Shelley Staples, CROW and MACAWS project leader, and two of her team members, Dr. Bruna Sommer-Farias and Dr. Nina Conrad; for their time to meet with us and give us access to their project materials and sharing their lessons learned with us; and Dr. Susan Conrad, leader of the Civil Engineering corpus project, for making available for us a number of helpful teaching materials derived from that corpus.

11 Zappa-Hollman, S., Ferreira, A. A., Perris, G. & Matte, M. L. (March 2022). Designing a local learner corpus for pedagogical applications and research. Paper presentation at the Virtual *TESOL Annual Convention*.

References

Blanchard, K. & Root, C. (2017). *Ready to write 3: From paragraph to essay* (4th Edition). Pearson.

Coffin, C. (1997). Constructing and giving value to the past: An investigation into secondary school history. In F. Christie & J. R. Martin (Eds.), *Genre and institutions: Social processes in the workplace and school* (pp. 196-230). Cassell

Conrad, S. (2017). A comparison of practitioner and student writing in civil engineering. *Journal of Engineering Education*, 106, 191-217. doi:10.1002/jee.20161.

Duff, P. (2008). *Case study research in applied linguistics*. Lawrence Erlbaum/Taylor & Francis.

Duff, P. A., Ferreira, A. A. & Zappa-Hollman, S. (2015). Putting (functional) grammar to work in content-based English for academic purposes instruction. In M. A. Christison, D. Christian, P. A. Duff, & N. Spada (Eds.), *Teaching and learning English grammar: Research findings and future directions: A festschrift for Betty Azar* (pp.139–158). Routledge.

Ferreira, A. A. (2020). Sociocultural development in the spectrum of concrete and abstract ideation, *Mind, Culture, and Activity*, 27(1), 50-69, doi:10.1080/10749039.2019.1686027

Ferreira, A. & Zappa-Hollman, S. (2019). Disciplinary registers in a first-year program. A view from the context of curriculum. *Language, Context and Text*, 1(1), 148-193. <https://doi.org/10.1075/langct.00007.fer>

Gardner, S. & Nesi, H. (2013). A classification of genre families in university student writing. *Applied Linguistics*, 34(1), 25-52. <https://doi.org/10.1093/applin/ams024>

Granger, S. (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. John Benjamins Publishing Company.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching*. John Benjamins, 13–32. <https://doi.org/10.1075/scl.33.04gra>

Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In Granger, S., Gilquin, G. & Meunier, F. (eds.) *The Cambridge handbook of learner corpus research*. Cambridge University Press, pp. 486-510.

Granger, S., Dupont, M., Meunier, F., Naets, H. & Paquot, M. (2020). *The International Corpus of Learner English*. Version 3. Presses universitaires de Louvain. Available at: <https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html>

Graves, R. (2017). Writing assignments across five academic programs. In R. Graves & T. Hyland (Eds.), *Writing assignments across university disciplines*. Trafford Publishing.

Graves, R. & Hyland, T. (Eds.) (2017). *Writing assignments across university disciplines*. Trafford Publishing.

Halliday, M. A. K. (1993). Towards a language-based theory of learning. *Linguistics and Education*, 5, 93–116. doi:10.1016/0898-5898(93)90026-7

Halliday, M. A. K. (1998). Things and relations: Regrammaticising experience as technical knowledge. In J. R. Martin & R. Veel (Eds.), *Reading science: Critical and functional perspectives*. Routledge. pp. 185–235.

Halliday, M. A. K. & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar* (4th ed.). Routledge.

Kang, J. (2016). A functional approach to the status of theme and textual development. *Theory and practice in language studies*, 6(5), 1053-1059. <http://dx.doi.org/10.17507/tpls.0605.20>

Matthiessen, C. M. I. M. (2006). Educating for advanced foreign language capacities: Exploring the meaning-making resources of languages systemic-functionally. In H. Byrnes (Ed.), *Advanced language learning: The contribution of Halliday and Vygotsky* (pp. 31–57). London, UK: Continuum.

Oshima, A. & Hogue, A. (2005) *Writing academic English* (4th ed.). Pearson-Longman.

Römer, U. & O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 6(2), 159-177. doi:10.3366/cor.2011.0011

Staples, S., Novikov, A., Picoral, A. & Sommer-Farias, B. (2019-). Multilingual Academic Corpus of Assignments – Writing & Speech (MACAWS). Available at <https://macaws.corporaproject.org>

Staples, S. & Dilger, B. (2018). *Corpus and repository of writing [learner corpus articulated with repository]*. Available at <https://crow.corporaproject.org>

Zappa-Hollman, S. (2018). Collaborations between EAP and disciplinary instructors: Factors and indicators of positive partnerships. *International Journal of Bilingual Education and Bilingualism*, 21(5), 591-606. doi:10.1080/13670050.2018.1491946

Zappa-Hollman, S. & Fox, J. (2021). Engaging in linguistically responsive instruction: Insights from a first-year university program for emergent multilingual learners. *TESOL Quarterly*, 55(4), 1081-1091. <https://doi.org/10.1002/tesq.3075>

Zappa-Hollman, S., Ferreira, A. A., Perris, G. & Matte, M. L. (March 2022). *Designing a local learner corpus for pedagogical applications and research*. Paper presentation at the Virtual TESOL Annual Convention.

Appendix A

Vantage Corpus of Student Texts Across Disciplines Project Survey

[Embedded institutional student consent form included here in original survey. The survey can be completed only after students provide informed consent]

Part 1 - Demographic information

Q1 What is your name? (as it appears in your UBC ID)

Q2 Please write down your preferred e-mail so that we can contact you:

Q3 Please confirm your email:

Q4 What Vantage Program are you in?

- Arts
- Science
- Engineering

Q5 How old are you?

- 17 to 19 years old
- 20 to 22 years old
- older than 22

Q6 What is your preferred gender?

- Male
- Female
- Other _____

Q7 What is (are) your native language(s)? You can select one or more, as it applies to you, to a maximum of three.

- Arabic
- Cantonese
- Farsi
- French
- German
- Hindi
- Indonesian
- Japanese
- Korean
- Malay
- Mandarin
- Mongolian
- Portuguese
- Russian
- Spanish
- Other _____

Q8 How many years of high-school education did you complete in **English**?

None

- 1
- 2
- 3
- 4
- More than 4

Q9 In what country did you receive your high-school diploma? (If none of the countries apply to you, please select Other at the end of the list.)

- Brazil
- Cambodia
- Canada
- Chile
- China
- Taiwan
- Ecuador
- Egypt
- France
- Germany
- Hong Kong
- India
- Indonesia
- Iran
- Japan
- Korea
- Macao
- Malaysia
- Mexico
- Mongolia
- Panama
- Russia
- Other

Q9a Other: In what country did you receive your high-school diploma?

End of Part 1 - Demographic Information (participants complete this once)

Part 2 - Assignment Information and upload

Q10 Would you like to upload another assignment?

- Yes

- No

Q11 Vantage College Corpus of Texts Across Disciplines Assignment information and uploading. Please, answer the following questions and then upload your assignment.

You will be prompted to answer the same questions for every assignment you upload.

Q12 Assignment upload:

Is this a single document?

- Yes
- No

Q12a You can upload only one document at a time. Please select another document and continue answering the questions.

Q13 Are you the only author of this assignment?

- Yes
- No

Q13a You can only submit an assignment completed by you only. Please select another assignment that you completed by yourself. Assignments completed together with your peers or classmates as part of pair/group work cannot be accepted.

Q14 Have you received a grade for this assignment?

- Yes
- No

Q14a You can only submit assignments that have been graded. Please select an assignment you have completed by yourself and for which you received a grade.

Q15 Course you completed this assignment for:

(NOTE: you can only upload assignments submitted only for the courses listed below)

Course

- VANT 140
- WRDS 150
- ASTU 204

Q16 What grade did you receive for this assignment?

- 90 - 100
- 70 - 89
- 60 - 69
- 50 - 59
- below 50
- Prefer not to answer

Q17 Upload your assignment:

Q18 If available, please upload the instructions you received to complete this assignment.

End of Part 2 - Assignment Information and upload

Part 3 - Interview Invitation

Q19 Thank you for uploading your assignment(s).

How easy or difficult was it to answer the questions and upload the assignment?

- Extremely easy
- Somewhat easy
- Neither easy nor difficult
- Somewhat difficult
- Extremely difficult

Q20 Would you be available to participate in a 30 minute interview to share your experience in this pilot project and to share with us information about the process of writing your assignment(s)?

For your participation in the interview you will receive a \$20 UBC Bookstore web gift card.

Do you want to participate?

- Yes, please send me more information about the interview.
- No

Q21 Is this the email you would like to be contacted at: [email entered by participant]?

- Yes
- No

Q22 Please, provide your preferred e-mail so that we can send you more information about the interview.

End of Part 3 - Interview Questions

Appendix B

VanCor Metadata Annotated Coding Sheet

Date coded: [yyyy/mm/dd]

Coder: [Name of person who coded]

Project: [name of LC project]

Assignment UID: [unique ID assigned to text being coded]

- Date submitted to instructor: [yyyy/mm/dd]
- Date submitted to VanCor: [yyyy/mm/dd; this is the date the assignment was uploaded by the participant to the Qualtrics survey]

Vantage Program: [select what applies]

Science

Engineering

Arts

Type of course: [Include here dropdown menu with list of courses from the corresponding program]

- EAP Writing course
 - LLED 200
 - LLED 201
- EAP disciplinary-linked course
 - VANT 140
- Other writing and communication course
 - ASTU 204
 - WRDS 150
- Disciplinary courses

Semester:

W1 [September-December]

W1-2 [September-April]

W2 [January-April]

S [May-July]

Course length in weeks: [include number of weeks]

Demographic Info:

- Age:
- Gender:
- Native language(s):

- Years of high-school education in English:
- Country received HS diploma:

Assignment

- Grade received:
- Percentage of final grade:
- Researcher's rating:
- Assignment **instructions** provided?
 - Yes
 - No
- Genre:
 - Instructor's label if provided: [this refers to the way the instructor called the genre of the assignment; e.g., annotated bibliography]
 - Student's label if provided: [this refers to how the student may have labeled the genre of the assignment; e.g., "in this discussion" - this is determined by looking at "clues" related to the overall structure of the text; e.g., "On the one hand...on the other hand..."]
 - Researcher's label: [use SFL-based classification]
- Is this assignment a component of a larger assignment? Yes/No
No
Yes: (link to genre of final assignment) (e.g. Results part of IMRD)

Length/# words:

Title:

Visuals included in the text? (e.g., figures, images, symbols, tables, graphs):

No

Yes

Completed In-class?

Yes

No

Completed out of class?

Timed

Not timed