RiCL Research in Corpus Linguistics

# The *Varieties for Specific Purposes dAtabase* (VESPA): Towards a multi-L1 and multi-register learner corpus of disciplinary writing

Magali Paquot[a] – Tove Larsson[b] – Hilde Hasselgård[c] – Signe O. Ebeling[c] – Damien De Meyere[a] – Larry Valentin[a] – Natalia J. Laso[d] – Isabel Verdaguer[d] – Sanne van Vuuren[e]

Université catholique de Louvain[a] / Belgium
Northern Arizona University[b] / USA
University of Oslo[c] / Norway
University of Barcelona[d] / Spain
Radboud University[e] / The Netherlands

**Abstract** – The *Varieties of English for Specific Purposes dAtabase* (VESPA first release) is the result of an international corpus compilation project that aims to address the lack of large-scale, open access, multi-L1, multi-discipline and multi-register learner corpora. This corpus report provides a detailed description of VESPA and illustrates possible uses of the corpus for register exploration of learner data. Specifically, it first offers an overview of the makeup of the corpus and the online interface that can be used to search and download the corpus. It then gives an illustrative example of a study where multi-dimensional analysis was used to investigate the relative importance of register vis-à-vis other factors in learner academic writing. In the concluding remarks, we identify priorities for future developments in the VESPA project, including the addition of more L1 components, more disciplines and more registers, as well as the compilation of a comparable corpus of native student writing.

**Keywords** – learner corpus; learner corpus research; English as a Foreign Language; academic writing, register variation; student writing

## 1. INTRODUCTION[1]

The main objectives of this corpus report are to provide a detailed description of the *Varieties of English for Specific Purposes dAtabase* (VESPA first release) and to illustrate how the corpus can be used to facilitate exploration of learner languages across registers

[1] We are most grateful to Paul Rayson (Lancaster University, UK) for giving us access to the CLAWS7 POS-tagger. We also thank Hubert Naets (UCLouvain, Belgium), main developer of the corpor@uclouvain.be platform, for his help at the initial stages of the project.

and different first-language (L1) backgrounds. As outlined below, corpora enabling large-scale, multi-L1, multi-discipline and multi-register investigations of learner language have previously not been available to researchers in the field. In making VESPA publicly available, we hope to help facilitate such studies, thus contributing one among many resources needed in order to provide a more accurate and nuanced picture of learner language.

Traditionally, the vast majority of written learner corpora available to the research community have included general argumentative or narrative texts produced by foreign language learners in the context of foreign/second language courses for general purposes (e.g. the *International Corpus of Learner English* (ICLE), 3rd edition, Granger *et al*. 2020). More recently, a number of learner corpora that comprise official language tests have also been released (e.g. *ETS Corpus of Non-Native Written English*, Blanchard *et al.* 2013; the *Open Cambridge Learner Corpus* 2017). By focusing almost exclusively on these contexts of use (and associated tasks), however, the field of learner corpus research has arguably developed a somewhat narrow perspective on what learner languages typically are. For example, overuse of first person pronouns, pragmatic inappropriateness and overstatements are linguistic features commonly reported in the literature to be typical of English as a Foreign Language (EFL) (e.g. Paquot 2010). This is somewhat problematic given that a growing body of research (e.g. Paquot *et al*. 2013; Larsson and Kaatari 2019) has noted that learners' use of many of these features (most particularly features related to writer-reader visibility) are often register-specific, thereby demonstrating the importance of including a broader range of registers in studies of learner language.

Further, in the context of English for Academic Purposes (EAP), the scope of registers analyzed to identify (i) typical characteristics of learner writing (development) and (ii) learners' difficulties remains overly restricted, meaning that the results of such studies often are of limited utility for EAP pedagogy. As stated by Biber *et al*. (2020: 49)

> university students are expected to produce a bewildering array of different registers, associated with the expectations of different disciplines, at different levels of study, and associated with the particular tasks required by their academic programs.

Therefore, there is a need for EAP researchers and practitioners to broaden their empirical basis. Corpora of EFL learner academic writing have been, or are being, compiled, but for different reasons, they are rarely available (Granger and Paquot 2013). Examples

include the *Corpus of Academic Learner English* (Callies and Zaytseya 2013) and the corpus of L2 disciplinary writing used in recent studies by Biber and colleagues (Staples *et al*. 2018; Biber *et al*. 2020). In addition, they often represent the writing of just one L1 population (e.g. German EFL learners in the *Aachen Corpus of Academic Writing*; Ströbel *et al*. 2020) or one register with a focus on dissertations (e.g. *Chinese Academic Written English Corpus*; Lee and Chen 2009). In that sense, the situation has not evolved much since Alsop and Nesi's (2009: 72) remark that discipline-specific student writing "has tended to be collected for individual scholarly purposes rather than as part of formal corpus-building projects."

While recently compiled open access corpora of academic writing such as the *British Academic Written English* corpus (BAWE; Nesi *et al*. 2008) and the *Michigan Corpus of Upper-level Student Papers* (MICUSP; Römer and O'Donnell 2011) include some texts by L2 writers, they were not compiled with a view to studying learner writing and/or learner writing development. Rather, the main objective of their collection is to investigate register and disciplinary differences in academic writing through a record of highly proficient university-level (mostly native-speaker) student writing. This means that only a limited number of learner texts per discipline or register are included; for example, there are only 39 EFL learner texts written in the field of linguistics in BAWE, with a variety of first languages represented (Bulgarian, Chinese, French, German, Greek, Italian, Japanese and Portuguese).

Given this lack of large-scale, open access, multi-L1, multi-discipline and multi-register corpora of learner academic writing, the VESPA learner corpus compilation project was initiated by Dr. Magali Paquot at the *Centre for English Corpus Linguistics* (CECL, UCLouvain, Belgium) with the aim to build a large collection of disciplinary writing by L2 English university students across registers and disciplines. Like other CECL corpora, VESPA is a corpus compilation project that involves collaborative work among several universities internationally. Partners have joined at different times and the corpus is still under compilation, with new components (e.g. new L1 backgrounds and more disciplines) continuously being added. The compilation process is described in detail in Section 2 together with an overview of the makeup of the corpus and the online interface.

While still work-in-progress, VESPA has already been used in a variety of studies to analyze linguistic features of EFL learners' academic writing in content courses (e.g.

Hasselgård 2014; Larsson 2019; Paquot 2019; Larsson *et al*. 2020), and to compare learners and native speakers' use of recurrent word combinations across disciplines (Ebeling and Hasselgård 2015). VESPA has also been used to complement data from other learner corpora such as ICLE: used together, the two learner corpora enable large-scale, multi-L1, multi-register explorations of learner data (Paquot *et al*. 2013; Larsson *et al*. 2021). With more subcorpora being added (especially subcorpora representing more disciplines) in the future, VESPA will also allow researchers to compare learner academic writing across registers and disciplines. In Section 3, we illustrate one of the many possible uses of VESPA by providing a brief overview of a recent study that made use of multi-dimensional analysis to investigate the relative importance of register vis-à-vis other factors in learner academic writing (Larsson *et al*. 2021). Finally, in Section 4, we make some concluding remarks.

## 2. VESPA: CORPUS COMPILATION, CORPUS PROCESSING AND ACCESS

In its current form (first release), VESPA comprises 941 texts (over 2 million words) produced by university students at the Bachelor's and Master's levels and collected by VESPA partners from five European universities (Radboud University, The Netherlands; UCLouvain, Belgium; University of Barcelona, Spain; University of Oslo, Norway; Uppsala University, Sweden), as shown in Table 1. The majority of the texts were written by students who have one of the official languages of the partner institutions (Dutch, French, Norwegian, Spanish, and Swedish, respectively) as their first language. Given the cultural diversity of some of the cities where the partner institutions are situated and the internationalization of higher education, however, 26 per cent of the collected texts across the various institutions represent academic writing by EFL learners with other L1 backgrounds than the official language of the respective institutions (examples of these other L1 backgrounds include Chinese, Czech, German, Greek, Italian, Polish, Russian, Turkish, and Vietnamese). 23 per cent of the students also report that they speak two languages or more at home.

| Institution | Main L1 language represented | Number of texts | Total number of words | Number of words per text (median [Q1 – Q3]) |
|---|---|---|---|---|
| Radboud University (The Netherlands) | Dutch | 118 | 310,099 | 2,616 [1,992 – 3,152] |
| UCLouvain (Belgium) | French | 154 | 648,483 | 4,072 [3,295 – 4,816] |
| University of Barcelona (Spain) | Spanish | 85 | 57,323 | 575 [525 – 755] |
| University of Oslo (Norway) | Norwegian | 515 | 772,964 | 1,180 [738 – 2,005] |
| Uppsala University (Sweden) | Swedish | 69 | 399,352 | 6,038 [2,894 – 7,634] |
| **Total** | | **941** | **2,188,221** | **1,809 [822 – 3,224]** |

Table 1: Corpus size per institution and main L1 language represented

With regard to the types of text included, VESPA comprises assignments that students submitted for course credit in disciplinary content courses. In that sense, the corpus answers repeated calls for greater ecological validity in L2 writing research (Polio 2017; Biber *et al*. 2020). As shown in Table 2, the large majority of the texts (79%) were collected in linguistic courses (taught by VESPA partners or colleagues in the same department) but some VESPA partners have also started compiling sub-corpora in literature and business communication.

| **Discipline** | **Number of texts** |
|---|---|
| Linguistics | 741 |
| Business communication | 126 |
| Literature | 74 |
| Total | 941 |

Table 2: Disciplines represented in VESPA

To classify the VESPA texts into register categories, we used the classification system from MICUSP (Römer and O'Donnell 2011: 170–171), which has two main advantages: the number of text categories is limited to seven, and each category comes with a set of defining linguistic features that can serve as simple guidelines. Table 3 provides an overview of the texts across the five register categories currently represented (critique/evaluation, proposal, report, research paper and response paper). This categorization is the result of an annotation procedure where each text was coded either using the register category identified by looking at the course requirements or, for the cases where we did not have access to the course requirements or could not obtain the information from the course instructor, texts were double coded by two VESPA partners (or a VESPA partner and a trained research assistant). Any disagreements were discussed and resolved with the VESPA coordinator. As shown in Table 3, the majority of texts (78%) fall into one of two categories: reports and research papers. However, given that texts were collected in different courses with different requirements at different institutions, the corpus is not balanced in terms of register by L1.

| Institution | Radboud University (The Netherlands) | UCLouvain (Belgium) | University of Oslo (Norway) | University of Barcelona (Spain) | Uppsala University (Sweden) | |
|---|---|---|---|---|---|---|
| *Main L1 represented* | *Dutch* | *French* | *Norwegian* | *Spanish* | *Swedish* | |
| **Registers** | | | | | | **Total** |
| Critique / evaluation | 5 | 3 | 129 | 0 | 0 | **137** |
| Proposal | 45 | 0 | 0 | 0 | 0 | **45** |
| Report | 26 | 36 | 268 | 85 | 0 | **415** |
| Research paper | 42 | 115 | 93 | 0 | 69 | **319** |
| Response paper | 0 | 0 | 25 | 0 | 0 | **25** |
| **Total** | **118** | **154** | **515** | **85** | **69** | **941** |

Table 3: Registers represented in VESPA

Table 4 provides information about the main rhetorical purpose of each register, its defining features and examples as detailed in Römer and O'Donnell (2011).

| Register | Rhetorical purpose | Defining features | Example |
|---|---|---|---|
| Critique/evaluation | Presents a positive or negative assessment of an outside source/project/text | - The text is driven by an in-depth assessment of a product/policy/procedure/text (although often interwoven with a description or observation of the product/policy/procedure/text)<br>- Gauges the effectiveness, validity, or usefulness of something<br>- Recommendations for improvement may be offered | Evaluation of business practices, problem-solution, literary critique, operations report |
| Proposal | Puts forth a research question, a theory or a model that the author feels should be explored in order to further the understanding of a given topic | - Formulates a research question or model, or proposes a potential study<br>- Usually does not collect or synthesize new data, but may include projected results; any collected data will be to support the proposal<br>- Justifies the need for data collection or data verification<br>- Critiques relevant literature and/or prior studies | Research proposal |
| Report | Describes the state or gives an account of a problem/issue/text, or describes the carrying out of a procedure (demonstrates the ability to gather data and summarize) | - Most space is devoted to description, rather than critical assessment<br>- Not driven by an original thesis or research question<br>- Author's opinion/evaluation may be present, but is not foregrounded and does not appear to drive the text | Lab report, literature review, article review, annotated bibliography, compare/contrast paper |
| Research paper | Presents original research in the field | - Entire text serves to answer a clearly stated research question<br>- Contains original data, or compiles existing data for the purpose of providing a new interpretation<br>- Structured into predictable sections (usually with subheadings)<br>- Includes most of the following: abstract, literature review, methods, results, discussion, conclusion | Research paper, replication study |
| Response paper | Short piece of writing responding to a given prompt or question, although prompt may not be explicit in the text | - Short in length (typically 1-2 pages)<br>- Style tends to be informal (e.g. expressions of emotional response; frequent references to mental processes, such as 'I was confused', 'I was surprised')<br>- May lack a formal introduction/'jumps right in' to content of paper, because author assumes reader's familiarity with the given topic (shared knowledge or in-group knowledge)<br>- The text provokes new questions for the author that may not be thoroughly answered | Solution to a homework problem, personal response to a text |

Table 4: VESPA text categories and definitions for text classification (adapted from MICUSP paper categories, Table 5 in Römer and O'Donnell 2011: 170–171)

The VESPA corpus compilation followed the same procedure across all institutions; this procedure aimed to maximize homogeneity of texts by applying the same inclusion criteria for all the texts across all institutions. First, we recruited students in specific content courses via their instructors.[2] The students filled out a questionnaire that is used to collect a set of learner and task variables (e.g. first language, level of study, number of years studying English at university, and content course for which the text was written) as well as a permission form. Both files are available in paper format and as an online survey. Second, the VESPA partner(s) at each institution collected the student work in electronic format, typically as Microsoft Word documents, and then annotated and processed the files with a series of tools developed or adapted for the project. These steps resulted in marked-up .xml files that are then ready for inclusion into VESPA. More specifically, following the procedure used in the BAWE corpus (Ebeling and Heuboeck 2007; Heuboeck *et al*. 2008), the texts were first processed using Word macros to annotate main sections (e.g. abstract, introduction), block quotes and so-called mentioned items (e.g. cited works, foreign words, linguistic examples). Next, the annotated texts were processed by means of Perl scripts to produce .xml files that include both the text and the metadata.[3] The complete corpus compilation procedure is described in the VESPA manual (Paquot *et al*. 2015).[4]

VESPA is available open access for non-profit educational and/or linguistic research purposes from the corpor@uclouvain.be platform, an online catalogue of corpora compiled at UCLouvain.[5] The platform can be used to search or download the corpus, in parts or in whole. Users first select texts by ticking variables of interest (e.g., all texts written in linguistics courses by French EFL learners) in the first three tabs of the 'Text selection' menu (Learner variables I, Learner variables II, and Task variables). Figure 1 shows the 'Task variables' page. The distribution of texts for each variable is dynamic; for example, in VESPA as a whole, there are more texts at the Bachelor's level than at the Master's level. However, if Radboud University is the only university that is

[2] Note that this is the main reason why each partner started with the collection of papers written in linguistic courses. Most of the time, VESPA partners were also the instructors for these courses and had direct access to the students and their writing.
[3] The Word macros and Perl scripts were developed by Alois Heuboeck (Reading University, UK); they are largely based on what was developed for the *British Academic Written English* (BAWE) corpus (cf. Ebeling and Heuboeck 2007; Heuboeck *et al*. 2008).
[4] The corpus collection guidelines and all associated material (student questionnaire, permission form, and Word macros) are available at https://tinyurl.com/VESPAguidelines.
[5] https://corpora.uclouvain.be/cecl/vespa/

ticked in the institution variable, the figures are recomputed for that particular institution, and we see that, in this subcorpus, the majority of texts were collected at the Master's level. As shown of Figure 2, the distribution of texts can also be explored graphically.

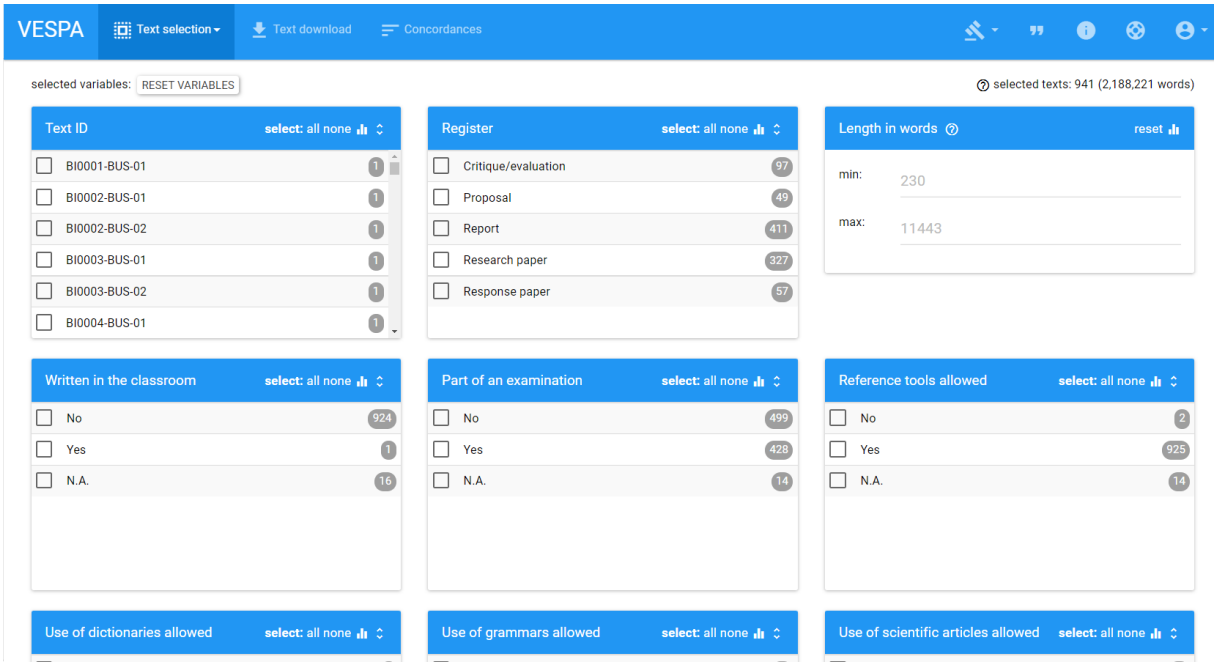


Figure 1: Selecting VESPA texts

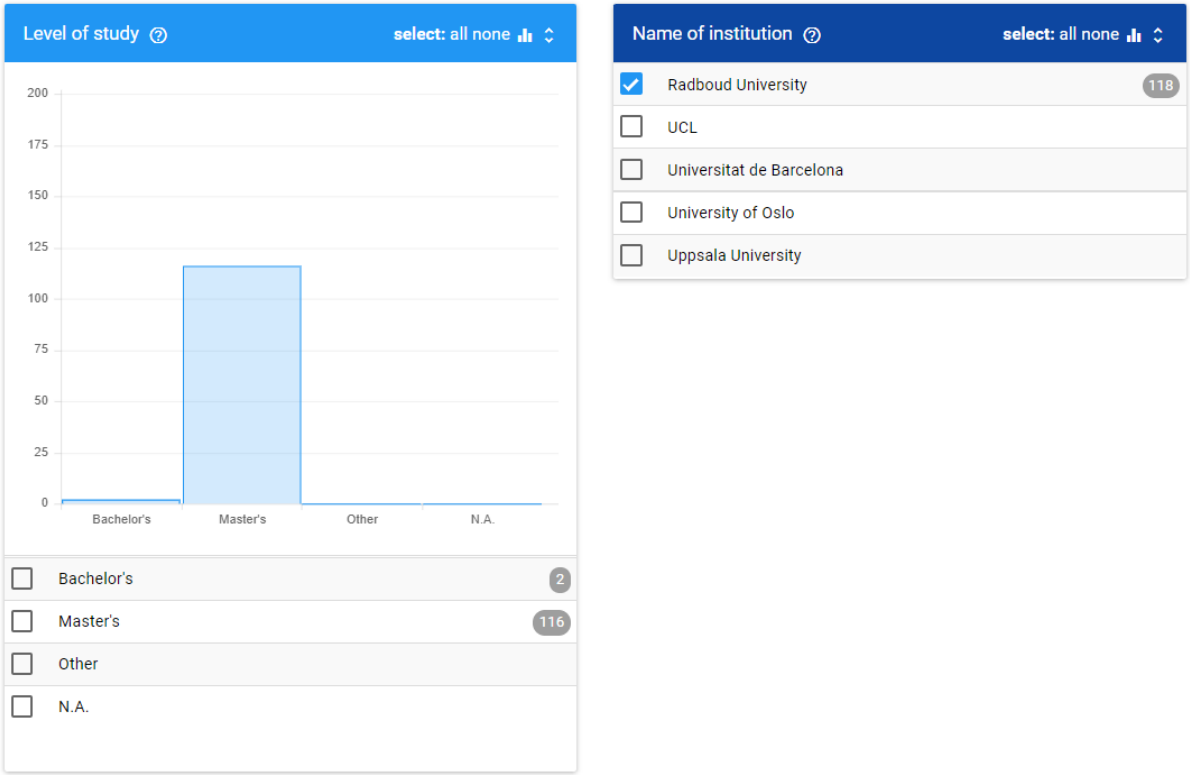


Figure 2: Exploring the corpus with the https://corpora.uclouvain.be/catalog/ platform

When a set of texts has been selected, the user can download it as a .zip file that will contain:

- A folder containing separate txt files for each text in the corpus (in UTF-8 format, no header);
- A file grouping all the texts in the corpus in a single txt file (in UTF-8 format);
- A database containing the learner profile information (metadata) for each text in the corpus in both .csv and .xlsx formats.

Alternatively, the selected texts can be explored online with a built-in concordancer that was initially developed for the third version of ICLE (Granger *et al*. 2020). One major improvement to the system is that it is configured to only search for linguistic items produced by EFL learners. Thus, if a user searches for the connector *however*, occurrences found in block quotes and mentioned items (see above) will not be retrieved.

All texts in VESPA are lemmatized and part-of-speech (POS) tagged with CLAWS7.[6] The concordance therefore makes it possible to search for word forms, lemmas, POS tags as well as combinations of word forms and lemmas with POS tags (see Part IV of Granger *et al*. 2020 for more details). Note, however, that the results of the automatic annotation were not manually checked and users of the corpor@uclouvain.be platform should check their accuracy when conducting a linguistic study that relies on lemma- and/or POS-based queries. Figure 3 shows the results of a search for the sequence *it* + modal verb + *be* + past participle in the whole corpus. Such concordances can then be exported in .xlsx or .csv format together with associated metadata, thus facilitating further analysis and treatment of the data outside the interface.

[6] https://ucrel.lancs.ac.uk/claws7tags.html

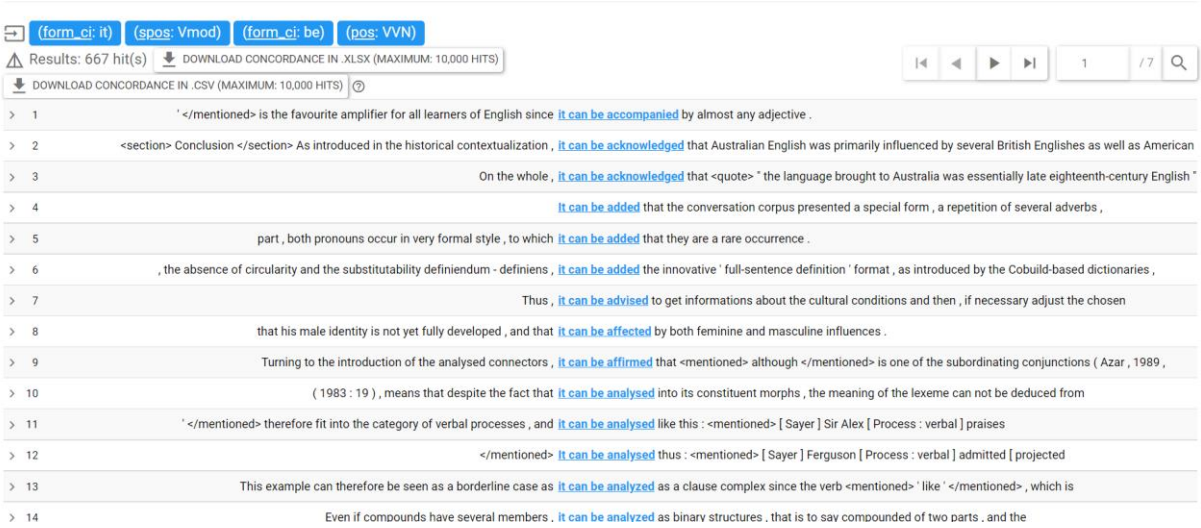


Figure 3: Searching VESPA with the corpor@uclouvain in-built concordancer

## 3. MAKING USE OF VESPA TO EXPLORE REGISTER VARIATION

As mentioned in Section 1, VESPA can be used for many different kinds of multi-L1 register comparisons, especially as a complement to the widely used ICLE (which almost exclusively includes argumentative essays). We will here illustrate this line of research by means of a recent study that made use of multi-dimensional (MD) analysis (Biber 1988) to examine learner and native-speaker student writing from two registers (argumentative essays and research papers) and published scientific articles, with the aim of investigating possible register effects in EFL learner writing. MD analysis is an approach used to describe and compare registers employing a wide range of linguistic co-occurrence patterns reduced to a few underlying 'dimensions' of variation that are then interpreted functionally (for a more detailed account of MD analysis, see Biber 1988, 1992). As such, the approach is ideally suited to investigate the extent to which features commonly attributed to EFL learner writing should be seen as more general characteristics of learner writing, as indicated in previous studies, or whether they may instead be prompted by (or at least moderated by) the register investigated. As shown in Table 5, the selection of corpora included in this study allowed for several different comparisons:

- Argumentative essays vs. research papers[7] vs. scientific articles (ICLE + LOCNESS vs. VESPA + BAWE + MICUSP vs. LOCRA)

[7] It is important to note that when the study reported on in Larsson *et al*. (2021) was conducted, the more detailed register categorization of VESPA texts had not been conducted yet. In that study, the term 'research paper' was used in a broader sense, as a superordinate category to refer to any piece of academic disciplinary

- Non-native vs. native speakers of English (ICLE + VESPA vs. LOCNESS + BAWE)
- L1 background (French, Spanish, Norwegian, Swedish and Dutch)

| Corpus | L1 | Register | Number of words | Number of texts |
|---|---|---|---|---|
| ICLE | French, Spanish, Norwegian, Swedish and Dutch | Argumentative essays | 708,541 | 1,073 |
| LOCNESS | English | Argumentative essays | 99,520 | 88 |
| VESPA | French, Spanish, Norwegian, Swedish and Dutch | Research papers in linguistics | 1,303,278 | 584 |
| BAWE | (British) English | Research papers in linguistics | 167,482 | 76 |
| MICUSP | (American) English | Research papers in linguistics | 313,785 | 34 |
| LOCRA | NA | Scientific articles in linguistics | 956,761 | 109 |
| **Total** | | | **3,549,367** | **1,964** |

Table 5: Overview of the corpora used in Larsson *et al*. (2021)

The results of the multi-dimensional analysis showed that the features investigated vary along two dimensions in the texts: 'Personal vs. topic-focused style' (Dimension 1) and 'Evaluative style vs. factual descriptions' (Dimension 2). While the study also reported certain differences across native vs. non-native status or L1 groups, the main differences were found between the registers, stressing its importance as a moderating variable. With both dimensions taken together, the novice writers' research papers (natives and non-natives) and the experts' scientific articles were found to be characterized by topic-focused and factual descriptions, the scientific articles significantly more so than the research papers. By contrast, the argumentative essays were shown to be personal and evaluative (L2 learners) or personal and topic-focused (English L1 students). Only very limited evidence was found to support claims made in previous studies about learner-specific characteristics such as a more involved style.

Larsson *et al*.'s (2021) results provide empirical evidence to support the increasingly more accepted view that "if we limit our investigations to argumentative writing only, the findings are likely to reflect that register and the results cannot (and should not) be used to make general claims about 'learner writing'" (Larsson *et al*. 2021: 254). The release of VESPA and its newly developed register classification will enable further explorations of learner (disciplinary) writing across more varied and specific

writing that provides analysis, interpretation, and/or argument based on independent research work. As such, the different register categories represented in VESPA are subsumed under this broader category (see Table 3).

registers than have often been the focus of previous research. With its focus on specialized registers in academic writing, VESPA can help answer (sometimes widely debated) questions such as (i) What are the main difficulties L2 writers face in an academic setting?; (ii) Are EFL learners' needs the same across disciplines and registers?; (iii) Does it make sense to provide general EAP courses?; and (iv) To what extent are L2 learners' needs the same as those of novice L1 students in an academic setting? (e.g. Gilquin *et al*. 2007; Römer 2009).

## 4. Conclusion

This corpus report has served to introduce VESPA and illustrate some of its many uses. While the corpus in its current form has already proven useful for describing linguistic features typical of specific types of disciplinary writing (mostly linguistics), and comparing learner features across registers, it is our belief that the following developments will make the corpus even more useful for the research community in the future. First, more partners have joined the project and corpora of disciplinary writing by Czech, Filipino and Turkish students are currently under development. Second, VESPA will soon also include comparable data in the discipline of linguistics by English-speaking L1 students. Third, we are also exploring avenues to collect data in other disciplines than linguistics, literature, and business.

It is our hope that the release of VESPA coupled with the publication of this corpus report will serve to inspire more research on learner languages across registers and disciplines.

## References

Alsop, Sian and Hilary Nesi. 2009. Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora* 4 /1: 71–83

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1992. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities* 26: 331–345.

Biber, Douglas, Randi Reppen, Shelley Staples and Jesse Egbert. 2020. Exploring the longitudinal development of grammatical complexity in the disciplinary writing of L2-English university students. *International Journal of Learner Corpus Research* 6/1: 38–71.

Blanchard, Daniel, Joel Tetreault, Derrick Higgins, Aoife Cahill and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013/2: i–15. https://doi.org/10.1002/j.2333-8504.2013.tb02331.x (29 September, 2021.)

Callies, Marcus and Ekaterina Zaytseva. 2013. The Corpus of Academic Learner English (CALE) – A new resource for the assessment of writing proficiency in the academic register. *Dutch Journal of Applied Linguistics* 2/1: 126–132.

Ebeling, Signe O. and Hilde Hasselgård. 2015. Learners' and native speakers' use of recurrent word-combinations across disciplines. In Ann-Kristin H. Gujord, Susan Nacey, Silje Ragnhildstveit eds. *Learner Corpus Research: LCR2013 Conference Proceedings* (Bergen Language and Linguistics Studies 6), 87–106.

Ebeling, Signe O. and Alois Heuboeck. 2007. Encoding document information in a corpus of student writing: The British Academic Written English Corpus. *Corpora* 2/2: 241–256.

Gilquin, Gaëtanelle, Sylviane Granger and Magali Paquot. 2007. Learner corpora: The missing link in EAP pedagogy. In Paul Thompson ed. *Corpus-based EAP Pedagogy. Special issue of the Journal of English for Academic Purposes* 6/4: 319–335.

Granger, Sylviane, Maïté Dupont, Fanny Meunier, Hubert Naets and Magali Paquot. 2020. *The International Corpus of Learner English* (version 3). Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, Sylviane and Magali Paquot. 2013. Language for specific purposes learner corpora. In Carol A. Chapelle ed. *The Encyclopedia of Applied Linguistics*. Oxford: Blackwell-Wiley.

Hasselgård, Hilde. 2014. *It*-clefts in English L1 and L2 academic writing. In Kristin Davidse, Caroline Gentens, Lobke Ghesquière and Lieven Vandelanotte eds. *Corpus Interrogation and Grammatical Patterns*. Amsterdam: John Benjamins, 295–320.

Heuboeck, Alois, Jasper Holmes and Hilary Nesi. 2008. The BAWE Corpus Manual. http://www.reading.ac.uk/AcaDepts/ll/app_ling/internal/bawe/BAWE.documentation.pdf (29 September, 2021.)

Larsson, Tove. 2019. Grammatical stance marking in student and expert production: Revisiting the informal-formal dichotomy. *Register Studies* 1/2: 243–268.

Larsson, Tove, Marcus Callies, Hilde Hasselgård, Natalia J. Laso, Magali Paquot, Sanne van Vuuren and Isabel Verdaguer. 2020. Adverb placement in EFL academic writing: Going beyond syntactic transfer. *International Journal of Corpus Linguistics* 25/2: 155–184.

Larsson, Tove and Henrik Kaatari. 2019. Extraposition in learner and expert writing: Exploring (in)formality and the impact of register. *International Journal of Learner Corpus Research* 5/1: 33–62.

Larsson, Tove, Magali Paquot and Douglas Biber. 2021. On the importance of register in learner writing: A multi-dimensional approach. In Elena Seoane and Douglas Biber eds. *Corpus-based Approaches to Register Variation*. Amsterdam: John Benjamins, 235–258.

Lee, David Y. W. and Sylvia Xiao Chen. 2009. Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing* 18/3: 149–165.

Nesi, Hilary, Sheena Gardner, Paul Thompson and Paul Wickens. 2008. *British Academic Written English Corpus*. Oxford Text Archive. http://hdl.handle.net/20.500.12024/2539

Open Cambridge Learner Corpus (v1). 2017. Distributed by Lexical Computing Limited on behalf of Cambridge University Press and Cambridge English Language Assessment.

Paquot, Magali. 2010. *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London: Continuum.

Paquot, Magali. 2019. The phraseological dimension in interlanguage complexity research. *Second Language Research* 35/1: 121–145.

Paquot, Magali, Hilde Hasselgård and Signe O. Ebeling. 2013. Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier eds. *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Louvain-la-Neuve: Presses universitaires de Louvain, 377–387.

Paquot, Magali, Signe O. Ebeling, Alois Heuboeck and Larry Valentin. 2015. *The VESPA Tagging Manual* (version 2.3). Louvain-la-Neuve: Centre for English Corpus Linguistics.

Polio, Charlene. 2017. Second language writing development: A research agenda. *Language Teaching* 50/2: 261–275.

Römer, Ute. 2009. English in academia: Does nativeness matter? *Anglistik: International Journal of English Studies* 20/2: 89–100.

Römer, Ute and Matthew Brook O'Donnell. 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.

Staples, Shelley, Douglas Biber and Randi Reppen. 2018. Using corpus-based register analysis to explore authenticity of high-stakes language exams: A register comparison of TOEFL iBT and disciplinary writing tasks. *The Modern Language Journal* 102/2: 310–332.

Ströbel, Marcus, Elma Kerz and Daniel Wiechmann. 2020. The relationship between first and second language writing: Investigating the effects of first language complexity on second language complexity in advanced stages of learning. *Language Learning* 70/3: 732–767.

*Corresponding author*
Magali Paquot
SSH/ILC
Collège Erasme
Place Blaise Pascal 1, bte L3.03.31
1348 Louvain-la-Neuve
Belgium
e-mail: magali.paquot@uclouvain.be

received: October 2021
accepted: June 2022