Data-driven learning in ESP university settings in Romania: multiple corpus consultation approaches for academic writing support

Mădălina Chitez[1] and Loredana Bercuci[2]

Abstract. Corpora are valuable technology-supported learning resources to be used by autonomous language learners or during teacher-guided lessons. This study explores the potential of corpus consultation approaches for the improvement of English for Specific Purposes (ESP) students' academic writing skills. We investigated the effects of three types of Data-Driven Learning (DDL) activities in a sample group of 29 first-year and second-year students majoring in Geography for Tourism at a Romanian university, consisting of writing tasks supported by: a Learner Corpus (LC), a Native-Speaker Corpus (NSC), and a Web-based Corpus (WBC). The research methodology involves the combination of quantitative and qualitative data, extracted from pre- and post-intervention corpus analyses, with the results of a learner-satisfaction questionnaire. The findings indicate a significant differentiation in the complexity of the lexico-grammatical features used by learners in consequent intervention stages and a better integration of L2-related academic writing strategies into their written productions. The study yields first conclusions on the integration of computer-processed language databases in DDL strategies for ESP learners in the Romanian university context.

Keywords: academic writing, corpus consultation, learner corpus, expert corpus, Romanian academic writing, DDL, English for specific purposes, Romanian learner corpus.

1. West University of Timişoara, Timişoara, Romania; madalina.chitez@e-uvt.ro
2. West University of Timişoara, Timişoara, Romania; loredana.bercuci@e-uvt.ro

1.   Introduction

Courses in ESP at the undergraduate level in Romania often have the aim of preparing students for their future profession, with the focus being on lexis and less frequently on writing. Typical ESP activities train students for a variety of real-life situations but often do not include academic communication and writing. Moreover, undergraduate students themselves are less motivated to become skilled in academic writing, perceiving it as difficult and unimportant for their careers. In reality, however, many Romanian undergraduate students pursue a master's degree, sometimes in English, and have trouble managing Anglo-American academic writing genre norms. The aim of this paper is to investigate whether certain DDL strategies (Boulton, 2017; Gilquin & Granger, 2010), such as multiple corpus consultation, can be used to teach lexis and academic writing concurrently.

2.   Method

The present study investigates the outcome of a pedagogical experiment in which course participants used three types of corpus-based DDL activities to (1) identify challenges related to common lexis use, discipline-specific jargon, and academic writing norms, and (2) find suitable solutions for their own academic writing difficulties. The experiment was conducted at a Romanian university, as part of an ESP course, at a geography department. The participants were 19 first-year (Common European Framework of Reference for languages – CEFR – level B1) and ten second-year (CEFR level B2) undergraduate students specializing in Geography for Tourism. The L1 of all participants is Romanian.

Each student was first asked to produce a short research essay on a set topic. The essays were compiled into a learner corpus, TourLRN. In the first session thereafter, the students were asked to compare, using LancsBox (Brezina, Timperley, & McEnery, 2018), the most frequent words used in their texts to the LOCNESS corpus, an English NSC. They identified the following words as subject to overuse: *the*, *of*, *that, you, people*. The students were then asked to rephrase parts of their essays, as much as possible, to use the identified words less.

In the second session, the students were introduced to the British National Corpus (BNC) and asked to select two problematic words of phrases in their texts written in the first session. Each student used the BNC to discover collocations containing the selected words, which were largely non-specialized terms. The students were asked to include the collocations in their texts (Figure 1).

Figure 1.   Methodology of corpus consultation intervention
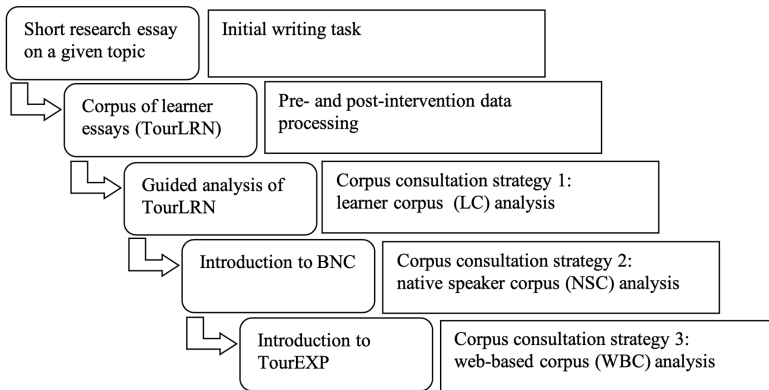
In the third session, the students used LancsBox (Brezina et al., 2018) to analyse an expert corpus, TourEXP, compiled by the researchers' team for the present study. The students were invited to identify discipline-specific terms and ngrams/collocates, as well as discipline-specific genre markers. They were asked to include at least three terms and four genre markers in their texts. The students were also introduced to the Whelk function in LancsBox and were encouraged to become familiar with the context of use for their chosen terms/phrases.

In all three sessions, students submitted un-revised versions of their texts and the teacher only pointed out group mistakes or linguistic inaccuracy patterns (e.g. overuse of *the*).

3.   Data

3.1.   Self-compiled corpora

For this study, we compiled two corpora: TourEXP and TourLRN. TourEXP is a web-based expert corpus made up of 155,521 tokens and was used solely for the purpose of in-class corpus consultation by the students. TourLRN is a learner corpus consisting of three sub-corpora: Batch 1 (pre-intervention texts, 8,176 tokens), Batch 2 (post-intervention texts, after the first corpus consultation session, 7,105 tokens), and Batch 3 (post-intervention texts, after the second corpus consultation session, 8,371 tokens). We performed a contrastive corpus analysis of the three versions of our students' texts.

3.2. Online questionnaire

At the end of the intervention study, the students were asked to fill in an online questionnaire in Romanian to gauge their perception of the utility of DDL techniques used in class. The questionnaire had 22 respondents.

4. Analysis

4.1. Corpus analysis

Not all students submitted all three versions of the essays. For the sake of accuracy, we compared the essays in three stages, namely Analysis 1 (A1): Batch 1 to Batch 2 (18x2 essays), Analysis 2 (A2): Batch 2 to Batch 3 (18x2 essays), and Analysis 3 (A3): Batch 2 to Batch 3 (23x2 essays).

4.1.1. Basic frequencies

As the teacher's personal observation was that the texts improved, in their last version, we looked at the fluctuation of highly frequent tokens in the students' ESP academic writing at different intervention stages (Figure 2).

Figure 2.   Fluctuation of most frequent tokens form one corpus consultation stage to another
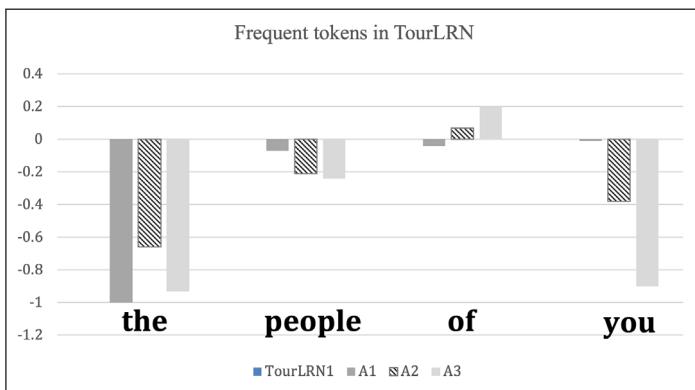
We noticed a decrease in the use of the definite article *the* (-1%) after the first corpus consultation exercise, then a slight increase (+ 0.34%) followed by a decrease in

Batch 3. A similar overall decrease pattern is noticed for *people* and *you*. On the other hand, students tended to use the preposition *of* more often than in their initial texts (TourLRN1) after being exposed to corpus data.

4.1.2. Ngram oscillation

We were also interested to see whether the discipline-specific lexical-grammatical constructions, i.e. collocations, were influenced by the use of corpora. Indeed, several typical collocations specific of the tourism sector were introduced in Batch 2 (*the tourism industry*) or Batch 3 (*in the tourism sector, travel insurance covers*). At the same time, several register makers oscillated toward formality: a decrease in the use of *a lot of* (by 0.05%), and disappearance, in Batch 3, of *a lot of people*, *a lot of things*. Several academic writing formulaic sequences also seemed to change the pattern of use from one text batch to another: *on the other hand* increased (by 0.2%) whereas *in conclusion* decreased (by 0.06%). However, the use of comparative transitions appeared to be challenging since appropriate rhetorical use was only observed in Batch 3 in a very limited number of texts.

Since the corpus size was rather small and created as a teaching exercise, we were content that most of our observations (correct use of *the*, diversification of the tourism terminology, or revision of informal style) were confirmed by absolute numbers. Percentages are used as indicators of use patterns.

4.2. Questionnaire analysis

Figure 3.   Questionnaire results – usefulness of corpora for academic writing
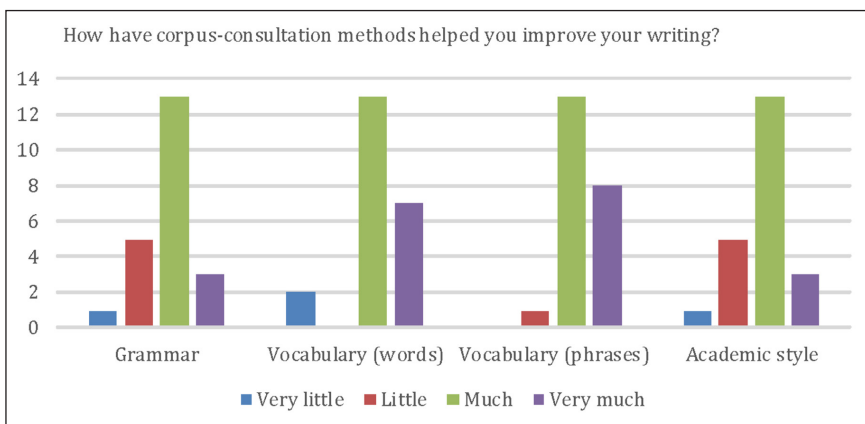
The questionnaire included questions such as: 'do you know what a corpus is?', 'which of the following types of corpora did you find most useful?', and 'how have the corpus-consultation methods helped you improve your writing?' (see Figure 3 above).

The results of the questionnaire were encouraging: all in all, all students admitted to have received information about the use of corpora for the first time during the evaluated course, they also considered the various methods of corpus consultation useful and they unanimously expressed their desire to learn more about corpora.

5.   Discussion and conclusions

Although quite experimental in design, the study was able to pinpoint areas of academic writing which can be supported by corpus linguistics in ESP courses. First, typical L1-L2 grammatical interference tendencies, such as overuse of the definite article *the* (also confirmed by Chitez, 2014) can be corrected by corpus consultation guided exercises, which involve comparison of students' own writing with expert writing. The tendency toward informality in ESP academic writing, observed in TourLRN, can also be corrected during corpus consultation training. ESP phraseology is imported and diversified as well at the end of all three corpus consultation stages. As for register appropriateness, we noticed an improvement at the ngram level, as typical academic writing markers were not only more frequently used but also better integrated in text. Due to the small size of the corpus, some of the fluctuations mentioned above are difficult to assess. However, modifying written assignments with the help of corpora was perceived as a positive experience by all the students in the intervention.

Our study shows that corpus consultation methods may be an effective way of stimulating inductive language (i.e. texts have been changed and improved according to observed corpus phenomena) and genre norm learning in ESP courses. Additionally, the motivational value for the students is confirmed by the questionnaire, thus offering encouraging prospects for further investigations.

6.   Acknowledgments

of the project ROGER (https://roger.projects.uvt.ro/), in progress at the West University of Timisoara, Romania, and funded by the Swiss National Science Foundation (program PROMYS).

References

Boulton A. (2017). Data-driven learning and language pedagogy. In S. Thorne & S. May (Eds), *Language, education and technology. Encyclopedia of language and education* (3rd ed.). Springer.

Brezina, V., Timperley, M., & McEnery, T. (2018). #LancsBox v. 4.x [software]. http://corpora.lancs.ac.uk/lancsbox

Chitez, M. (2014). *Learner corpus profiles: the case of Romanian learner English*. Peter Lang.

Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching? In A. O'Keeffe & M. McCarthy (Eds), *The Routledge handbook of corpus linguistics* (pp. 359-370). Routledge. https://doi.org/10.4324/9780203856949.ch26

CALL and complexity – short papers from EUROCALL 2019
Edited by Fanny Meunier, Julie Van de Vyver, Linda Bradley, and Sylvie Thouësny

Publication date: 2019/12/09