How to write good academic papers: using the EXPRES corpus to extract expert writing linguistic patterns

Madalina Chitez, Valentina Carina Mureșan, Roxana Rogobete

West University of Timisoara, Romania

Abstract

Successful academics and professionals need to master linguistic and text-rhetoric skills that have the potential to make disciplinary knowledge accessible to all sectors of society [1]. Some scholars define writing as "an important feature of a discipline's identity" [2], which foregrounds content-related language and linguistic challenges in discipline-specific academic discourse. Academic writing support turns, under the circumstances, into a vital component of the disciplinary expertise acquisition and sharing processes. The EXPRES corpus (Corpus of Expert Writing in Romanian and English) [3] is such a digital support tool: the corpus is composed of research articles in the disciplines, compiled using online resources. The profile of the corpus is "expert writing corpus" since all articles have been published in peer-reviewed journals. The EXPRES corpus has a dedicated corpus query platform which allows for the search and extraction of desired linguistic items (words, phrases, patterns, n-grams) and statistical visualizations of data. Starting from the assumption that "good writing" [4] can be linguistically stimulated, we used corpus specific research methods to extract linguistic patterns that can be useful in the process of academic paper writing. The extracted patterns are distributed into discipline specific patterns and general academic writing patterns. At the same time, as EXPRES is a Romanian-English bilingual corpus, all the linguistic outcomes also have language specific distributions and descriptions. The paper exemplifies and highlights the potential of using expert corpora for improving academic paper writing.

Keywords: *good writing, academic paper writing, EXPRES corpus, expert academic writing corpus*

1. Introduction

Nowadays, researchers all over the world are challenged to "publish or perish" [5]. The challenge becomes even greater considering the pressure of publishing in internationally recognized journals and volumes, whose publication language is English. Researchers and professionals in the disciplines are often hindered in their endeavours to disseminate science because of insufficient academic writing skills. In order to become expert writers, domain specialists go through various learning stages, sometimes referred to as mediation in sociocultural [6], be it formal (i.e. formal education) or informal (i.e. real-life experiences), explicit (i.e. institutional) or implicit (i.e. work-related) [7]. Academic (and professional) writing is, in this context, a key component of learning development: writing competence intersects with the co-construction of knowledge within professional communities [8].

In this context, the existence of an expert academic writing corpus, such as EXPRES - *Corpus of Expert Writing in Romanian and English* [9], dedicated to a specific language group (i.e. Romanian and English L2), can offer researchers the much needed linguistic instruments that would assist them in the process of writing research articles. The corpus has been compiled in the framework of the project DACRE - *Discipline-specific expert academic writing in Romanian and English: corpus-based contrastive analysis models*, conducted at the West University of Timisoara, Romania [1]. The project aims at identifying research-based practical solutions for writing in any of the two languages by pointing out academic writing differences, tendencies and specifics of the target discipline and the language used. In this paper, we exemplify the use of the EXPRES corpus data for linguistic support during the academic writing process.

2. Expert academic writing: language level perspectives

2.1 Academic Writing in English L1

Using corpus-based research to create guidelines for novice and experienced writers alike has a tradition of over two decades. Studies, such as [10], report the use of corpora to categorize and analyze linguistic phenomena in academic languages. Such studies used three main types of corpora to support their research:

(a) general corpora, such as the British National Corpus (BNC) or the Corpus of Contemporary American English (COCA); b) learner corpora, such as the British Academic Written English Corpus (BAWE), Michigan Corpus of Upper-Level Student Papers (MICUSP), Romanian Corpus of Academic Genres (ROGER), or (c) smaller, self-compiled case-study corpora, e.g. [11]. When it comes to discipline-specific writing, these types of corpora provide only limited resources for in-depth studies and applicability due to either their focus on general language use (no disciplines), the inclusion of texts representing mixed levels of expertise (non-experts) or restriction of public access. Only a few corpora, to our knowledge, focus exclusively on expert discipline-specific language.

2.2 Corpus-based academic writing in English L2

Discipline-specific corpora are a valuable tool for teaching English for Specific Purposes in general, and English for Academic Purposes in particular. First, they are resources with the help of which discipline-specific teaching materials can be created. For instance, a variety of studies have examined different linguistic aspects of discipline-specific academic texts, such as frequent lexical bundles, noun pre-modification, or register patterns. Such studies have been used as the basis for teaching materials, including online phrase banks, such as the Manchester Academic Phrasebank [12]. Furthermore, other studies have shown that corpora are effective tools when used in class for enhanced data-driven learning [11] or to increase student learning motivation [13].

2.3 Corpus-based academic writing in Romanian

As for Romanian corpora, except for several NLP-supporting corpora compiled at the RACAI Institute e.g. CoRoLa [14], there have been other, rather few, research initiatives that either use small case-study general corpora for translation studies [15], or have compiled corpora that focus on some professional fields, such as health, e.g. the Romanian Medical Corpus – MoNERo [16]. Corpora available on digital platforms are mostly limited to their presentation (in annotated or non-annotated format) or they are linked with simple search interfaces, but they are not used as practical user-friendly or further-research resources. As for corpus-based academic writing studies, the only existing extensive corpora are the Romanian Corpus of Learner English (RoCLE) [17] and the bilingual corpus Romanian Corpus of Academic Genres (ROGER) [18].

3. EXPRES corpus

3.1 Data

The EXPRES corpus (*Corpus of Expert Writing in Romanian and English)* [9] has been initially designed and compiled in four disciplines, but, since the corpus is dynamic, other disciplines can be added. The corpus consists of expert scientific writing texts, in two languages, Romanian and English, distributed into three distinct language varieties: Romanian L1, English L1 and English L2 texts (produced by Romanian speakers). The genre selected for inclusion in the corpus is the research article (RA), which, among academic genres, holds a central place, as it is the main form of scientific communication [19].

The size and structure of the corpus aims at ensuring balance and representativeness of the corpus data (see Table 1). Data collection is now in its final stages, with a final release of the corpus in July 2022. In order to collect data for the EXPRES corpus, various types of challenges had to be overcome, such as, for example, the availability of data for corpus extraction [3]. Free access to the corpus data is offered via a corpus query platform (under construction) to be accessible in July 2022.

Discipline	Sub-corpus (language-specific)	No of RAs / sub-corpus	Average no of words / RA	Approx. no. of words / set
Linguistics	RO-L1 / EN-L1 / EN-L2	200	3,000 w.	600,000 w. x 3
Economics	RO-L1 / EN-L1 / EN-L2	200	3,000 w.	600,000 w. x 3
Political Sciences	RO-L1 / EN-L1 / EN-L2	200	3,000 w.	600,000 w. x 3
Information Technology	RO-L1 / EN-L1 / EN-L2	200	3,000 w.	600,000 w. x 3
			Total EXPRES	**7,200,000 w.**

Table 1. Estimated size of the EXPRES corpus

3.2 Corpus query platform

Academic writing support functioning as a free-access database of authentic expert academic discourse (i.e. searchable academic writing corpus) is a relatively underexplored territory. The EXPRES corpus support platform is built in a similar manner with the ROGER corpus support platform [18]. The platform is designed to be a user-friendly tool where students, teachers, researchers and professionals can perform need-driven searches and explorations for practical use, teaching or research. The user interface displays the main platform parameters, which are automatically

processed depending on the amount of data administered by the platform: number of disciplines, languages, number of words, and number of characters. The rubrics of the platform are: Home, About, Corpus Documentation, Tutorials, Research and Statistics. The main functionality of the platform is signalled by the button "Sign in to search" where the user can create a free account. The user can then start searching the corpus for:

a. *Simple searches*: words (e.g. *paper*, morphological variations (e.g. paper* for <paper> and <papers>), string of multiple words, e.g. *in this paper*;

b. *Concordance lists*: each search is displayed as a concordance list (i.e. word in context) and can be viewed in more restricted or more general contexts (number of words);

c. *Data filtering*: all searchers can be filtered according to several criteria: language (Romanian, English L1 and English L2), discipline, journal visibility (low versus high visibility).

d. *N-gram lists* (i.e. frequent co-occurrences of more than 2 words); the first 100 n-grams are displayed in the user interface and the list can be downloaded in **.xslx* format.

3.3 Expert writing linguistic patterns

Analysing the two sub-corpora compiled within the DACRE project - Romanian-L1 and English-L2, we selected for this study a limited, but a comparable sample. Using digital instruments designed for corpus linguistics (such as Sketch Engine), an analysis of the most frequent N-Grams (sets of co-occurring words) is relevant in order to see discursive patterns present in research articles. Previous studies have shown that there is a dependency between the use of "fixed or semi-fixed multi-word sequences (MWSs)" such as N-Grams and the quality of producing written text [20]. Thus, we selected the first 10 tri-grams and four-grams from the two sub-corpora. The results are presented in Figure 1:

Romanian-L1	Relative frequency	English-L2	Relative frequency 2
în ceea ce privește	251.04	in order to	473.08
din punct de vedere	222.95	the fact that	338.25
pe de altă parte	195.44	as well as	316.18
pe de o parte	120.36	the case of	297.25
o serie de	214.36	one of the	284.64
cu privire la	222.95	in terms of	268.87
în funcție de	204.04	on the other	229.44
în timp ce	200.6	in the case of	211.31
în acest sens	163.35	part of the	205
cum ar fi	155.32	on the other hand	198.69

Fig. 1: The first 10 3-gram and 4-gram in RO-L1 and EN-L2

Another focus in improving academic writing is focused on formulaic sequences which can be used to create a clearer structure of the paper. In Romanian research articles (RAS), designing the introductory section rarely uses the CARS model – on account of a rather low level of explicit academic writing skill instruction, as well as a rather inconsistent set of article guidelines in Romanian journals, which lead to a greater diversity in the manner of writing, as shown in Figure 2. Phrases such as "The present article represents a…", "This article proposes a/an…", "The present approach involves a contrastive analysis…" alternate with a general description of the research area:

Left	Kwic	Right
	Introducere	Problema convergenței a intrat de mult timp în sfera de preocupări a economiștilor. </s><s> Aceștia au analizat dinamica
pe profit. </s><s> Cuvinte-cheie: Monedă virtuală, Bitcoin, contabilizare și raportare, impozitare Clasificare JEL: M41	INTRODUCERE	Începem prin definirea monedei virtuale (VC de la virtual currency). </s><s> Spre deosebire de monedele de modă tradiționale
inovatoare care să răspundă unui astfel de tip de expunere. </s><s> Cuvinte cheie: catastrofă, daune, asigurări obligatorii	Introducere	Din toate cercetările efectuate până în prezent de diferite entități abilitate rezultă că pierderile economice că

Fig. 2. Concordances of "introduction" (EXPRES-RO-L1 sub-corpus)

However, corpus tools that identify concordance lists of KWIK can be used in order to extract "good writing" examples, for instance in stating the aim or niche of the paper or depicting the study's limitations: "Since this study has an exploratory character", "For this specific study, we will limit the analysis to…", "Through this study, we join such analytical approaches…" etc.

25% din portofoliul fondurilor de pensii în active externe. </s><s> În acest	studiu	autorii evaluează, mai întâi, șapte modele diferite de optimizare pentru a
e studii lingvistice publicate de- a lungul timpului. 0.2. </s><s> Prin acest	studiu	, ne alăturăm și noi unor asemenea demersuri analitice, propunându-ne
ni o au biruit. 4.11. </s><s> Strâns legate de subiectul investigat în acest	studiu	sunt atât selectarea din fondul lexical medieval românesc a unor adjecti
oncretizat în ghidurile turistice ce constituie corpusul propus pentru acest	studiu	, atestă pertinența acestei constatări. </s><s> Prin exploatarea cu mijloa
or sintactice, al unor expresii și chiar al punctuației. </s><s> Pentru acest	studiu	, ne vom limita la identificarea, inventarierea și clasificarea adjectivelor e

Fig. 3. Concordances of "study" (EXPRES-RO-L1 sub-corpus)

A comparison with the English-L2 corpus shows that, surprisingly, there is a higher lexical variety in articles written by Romanian scholars in English (examples shown in Figure 4) - our hypothesis is that, to gain a higher degree of visibility, Romanian researchers tend to prefer English and are therefore aware of AW guidelines.

Left	Kwi	Right
From this perspective, we	aim	to reconsider Manfred's intention to translate the
The present paper	aims	to reconsider our approaches to the suppositio theory
All these efforts	aim	to try to answer the question: "What is the theory of suppositio and for what purpose was it made?" </s><s> .
The	aim	of the present study was to argue that the text Dissensiones inter viam antiquam et viam modernam is included in the via

Fig. 4. Concordances of “aim” (EXPRES-EN-L2 sub-corpus)

4. Discussion and conclusion

The educational value of EXPRESS is two-fold since it may serve both as a self-learning tool for professionals aiming to perfect their writing style and also open up new possibilities for activity design and vocabulary instruction in the classroom. Data-driven learning is a concept that is already familiar to language teachers nowadays and benefits such as making use of concordancing tools for language learning have already been explored since the 1990s [21], [22], but also fully explored nowadays [23]. In addition to providing greater reliability for contextualizing academic lexical structures, its importance for Romanian users also emerges from a lack of sufficient reliable instruments to facilitate the use of “academic vocabulary”, or check the accuracy of lexical structures in use (no dictionary of Romanian collocations). Thus the different types of searches within EXPRESS corpus will not just reflect lexical preferences and occurrences, but also structural combinations for the accurate usage of certain items in context (such as n-gram searches). This may further lead to a clearer understanding of the particularities of academic writing in Romanian or to the creation of resources such as academic vocabulary lists for novice users who write their texts in either Romanian or English L2.

Acknowledgement

https://dacre.projects.uvt.ro/

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number 158 / 2021, within PNCDI III, awarded to Dr Habil Madalina Chitez (PI), from the West University of Timisoara, Romania, for the project DACRE (*Discipline-specific expert academic writing in Romanian and English: corpus-based contrastive analysis models*, 2021-2022).

References

[1] Homepage. DACRE project webpage. Accessed 03 March 2022, https://dacre.projects.uvt.ro/?lang=en

[2] Costley, T.; Flowerdew, J. “Introduction”. In J. Flowerdew & T. Costley (eds.), *Discipline-specific writing. Theory and practice*, 2017, Routledge, pp. 1-11.

[3] Rogobete, R.; Chitez, M.; Mureșan, V.; Damian, B.; Duciuc, A.; Gherasim, C.; Bucur, A.-M. “Challenges in compiling expert corpora for academic writing support”. *Conference Proceedings, 11th International Conference the Future of Education*, Virtual Edition, Florence, Italy, Filodiritto Editore, 2021, pp. 409-414.

[4] Băniceru, C.; Tucan, D. “Perceptions about “Good Writing” and “Writing Competences” in Romanian academic writing practices: A questionnaire study”. In M. Chitez, C. Doroholschi, O., Kruse, Ł Salski,, & D. Tucan (eds.), *University Writing in Central and Eastern Europe: Tradition, Transition, and Innovation*, Springer, 2018, pp. 103-112.

[5] Moosa, I. A. *Publish or Perish: Perceived Benefits versus Unintended Consequences*. Cheltenham, Edward Elgar Publishing, 2018.

[6] Edwards, A. *Being an Expert Professional Practitioner: the Relational Turn in Expertise*. Netherlands, Springer, 2010.

[7] Wertsch, J. V. “Mediation”. In Daniels, M., Cole, M., and Wertsch, J. V. (Eds.), *The Cambridge companion to Vygotsky*, 2007, New York, Cambridge University Press, pp. 178-192.

[8] Sokolik, M. “Academic writing in MOOC environments: Challenges and rewards”. In Martín-Monje, E., Elorza, I., and Riaza, B. G. (Eds), *Technology-enhanced language learning for*

specialized domains: Practical applications and mobility, 2016, New York, Routledge. pp 165-176.

[9] Chitez, M., Rogobete, R., Muresan, V., and Dinca, A., *Corpus of Expert Writing in Romanian and English* (EXPRES). West University of Timisoara. Available at: https://dacre.projects.uvt.ro/research/?lang=en.

[10] Römer, U.; Cortes, V.; and Friginal, E. (Eds.). *Advances in Corpus-Based Research on Academic Writing: Effects of Discipline, Register, and Writer Expertise*. Amsterdam/Philadelphia, John Benjamins Publishing Company, 2020.

[11] Cotos, E. "Enhancing Writing Pedagogy with Learner Corpus Data". *ReCALL*, 26(2), 2014, pp. 202-224.

[12] Davis, M., and Morley, J. "Facilitating learning about academic phraseology: teaching activities for student writers". *Journal of Learning Development in Higher Education*, 2018, pp. 1-17. Available at: https://journal.aldinhe.ac.uk/index.php/jldhe/article/view/468.

[13] Chitez, M.; Bercuci, L. "Data-driven learning in ESP university settings in Romania: multiple corpus consultation approaches for academic writing support". In Meunier, F., Van de Vyver, J.; Bradley, L. & Thouësny, S. (Eds). *CALL and complexity – short papers from EUROCALL*. Research-publishing.net, 2019, pp. 75-81.

[14] Tufiș, D.; Ion, R.; Ceaușu A. and Stefănescu, D. "RACAI'S Linguistic Web Services". In *Proceedings of The 6th Language Resources and Evaluation Conference*, 2008, pp. 28-30.

[15] Popescu, T. (2013). "A Corpus-based Approach to Translation Error Analysis. A Case-study of Romanian EFL Learners". *Procedia - Social and Behavioral Sciences*, 83, 2013, pp. 242-247.

[16] Mitrofan, M, Barbu Mititelu V. and Mitrofan, G. "MoNERo: a Biomedical Gold Standard Corpus for the Romanian Language". *In Proceedings of the BioNLP 2019 workshop*, Florence, Italy, 2019, pp. 71-79. Available at: https://www.aclweb.org/anthology/W19-5008.pdf.

[17] Chitez, M. *Learner corpus profiles: the case of Romanian learner English*. Linguistic In-sights Series (Series Editor: Maurizio Gotti). Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien, Peter Lang, 2014.

[18] Chitez, M.; Bercuci, L.; Dincă, A.; Rogobete, R., & Csürös, K. *Corpus of Romanian Academic Genres* (ROGER). West University of Timisoara, 2021. Available at https://roger-corpus.org/.

[19] Swales, J. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, 1990.

[20] Garner, J.; Crossley, S.; & Kyle, K. "N-gram measures and L2 writing proficiency". *System*, vol. 80, pp. 176-187. https://doi.org/10.1016/j.system.2018.12.001. Available at: https://www.sciencedirect.com/science/article/pii/S0346251X1830201X.

[21] Aston, G. "Corpora in language pedagogy: Matching theory and practice". In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H.G. Widdowson*. Oxford, UK: Oxford University Press, 1995.

[22] Tribble, C. & Jones, G. *Concordances in the classroom: A resource book for teachers*. Essex, U.K.: Longman, 1990.

[23] Conrad, S. "Corpus linguistics and L2 teaching". In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers, 2005, pp. 393-409.