Challenges in Compiling Expert Corpora for Academic Writing Support

Roxana Rogobete[1], Mădălina Chitez[2], Valentina Mureșan[3], Bogdan Damian[4], Adrian Duciuc[5], Claudiu Gherasim[6], Ana-Maria Bucur[7]

West University of Timișoara, Romania[1, 2, 3, 4, 5, 6, 7]
University of Bucharest, Romania[7]

Abstract

Since most of the academic articles relevant for many disciplines are to be found in English, it is important to understand the linguistic challenges of academic publishing in English L2 in contrast with the mother tongue academic writing specifics. The present paper explores a series of challenges faced in the attempt to build expert corpora for academic writing research and teaching. Particularly, the study reports on the construction of the DACRE corpus, an expert bilingual comparable corpus consisting of discipline-specific peer-reviewed scientific articles. The corpus should facilitate the extraction of the salient linguistic and rhetorical features specific for each selected discipline (Linguistics, IT, Political Sciences, Economics) and language variety (Romanian, English L1 and L2). At the initial stage of the corpus compilation process, when assessing the linguistic resources to be included in the corpus, a multitude of challenges emerges. For example, the linguistic level of these resources is not consistent. Other difficulties we encountered were the data availability (open sources or subscription-based), lack of recent resources for certain corpus batches, "multi-authorship" in determining L1 texts, and, most important, legal aspects (i.e. copyright). By describing, comparing and analysing data collection obstacles, we propose a model for expert corpus building in English vs low-resource languages such as Romanian.

Keywords: *English vs Romanian academic writing, bilingual expert corpora, discipline-specific writing, Romanian expert corpus, DACRE corpus*

1. Expert writing in English L2 versus expert writing in the mother tongue

Scholars worldwide use English as the main *academic lingua franca* [1], [2]. The internationalisation path has also been assumed by the Romanian academic community in its endeavour to gain international recognition and impact. First, Romanian researchers publish in English in order to increase their visibility (either in national journals or in international periodicals). At the same time, English is commonly used as a medium of instruction in many different professional domains, so the Romanian universities have adapted rapidly to this development and discipline-specific English is part of numerous study programmes. Several fields (Linguistics, Economics, IT, Political Sciences) remain among the most frequent HE study programmes in Romania using EMI. It seems that the higher the degree of internationalisation of the domain, the greater the necessity to access and address the international research community in English, which requires proficient academic writing skills in English L2 on the part of the "academic writer". However, the same writer also needs to understand the differences in expectations regarding writing in English versus writing in his/her mother tongue in order to be successful in his/her disciplinary dissemination attempts.

2. Why expert corpora and what are they?

But how do scholars acquire/improve their academic writing skills? In general, there is a theoretical gap in terms of identifying linguistic patterns across field-specific academic texts. The Romanian writing cultures, for example, are scarcely researched [3], both in Romanian L1 and English L2, especially from a data-intensive perspective. As is the case with other languages, for the Romanian context, there is little research-based academic writing practice: scholars compensate for this by using observational models delivered by textbooks/classroom research papers. Systematically, at a larger scale, such practice can be supported and complemented by the use of expert writing corpora. Broadly defined, expert corpora are collections of texts that have been qualitatively validated, according to certain criteria, to be used for the extraction of linguistic data that serve as models of language use (see also [4]). Most expert corpora are L1 corpora, written in the user's mother tongue; however, data in certain corpora have to be pre-selected in case the writing is not guaranteed to be "expert" (e.g. written texts by student learners that have received poor grades). While there are

several corpora of expert English L1, for low-resource languages such as Romanian, there are limited instruments (Table 1).

Expert corpora in English L1	Expert corpora in other languages	Expert corpora in Romanian (part of specialised corpora)
COCA, BNC, MICUSP, BAWE	DWDS, KORP, CoRoLa, PAROLE	BioRo, ROMBAC
Other info: https://warwick.ac.uk/fac/soc/al/repository/staff/harrisontilly/corpora-for-workshop/	Other info: https://www.clarin.eu/portal/	Other info: https://www.sketchengine.eu/corpora-and-languages/romanian-text-corpora/; https://www.racai.ro/tools/text/

Table 1: Expert corpora – examples

Additionally, expert corpora in L2 writing need to be compiled according to clearly predefined criteria (Fig.1). Nonetheless, the accessibility of such corpora and resources is limited, either because access is licence-based (e.g. COCA) or they overlap with specialised corpora, with only subsets of data being "expert".

(a) professional writing;

(b) peer-reviewed published texts;

(c) market validated resources (e.g. Wikipedia or similar blog texts, high quality newspapers).

Fig.1: Criteria for expert corpora in L2

3. DACRE project

In order to compensate for the lack of research-based support for expert academic writing, DACRE (*Discipline-specific expert academic writing in Romanian and English: corpus-based contrastive analysis models*) has been initiated at the West University of Timisoara, Romania. The project includes the creation of a bilingual comparable corpus, consisting of peer-reviewed scientific articles from different disciplines: Linguistics/Political Sciences/Economics/IT. DACRE aims to popularise the use of corpora in HE and research-based practice and to create digital instruments, methodological analysis models useful to the national/international language-related research community. The intention is to facilitate the extraction of salient linguistic/rhetorical features specific to each discipline and each language variety (Romanian/English L1/English L2).

4. Challenges in building expert corpora

When assessing the linguistic resources to be included in the expert corpus DACRE, a multitude of challenges emerges:

4.1 Language related challenges

At the level of academic text selection a few caveats were identified. From the three types of languages targeted by the study the category of expert writing produced in Romanian L1 revealed a numerical imbalance between the academic writing samples from different fields (see Section 4.3).
Furthermore, in the case of English as L1, the main limitation concerns identifying the appurtenance of the author(s) to a L1 community. Although in ELT literature there are references to L1 versus L2 writing, this is discussed from the perspective of language learning and writing is seen as an indicator of linguistic proficiency, rather than reflecting the reliability of an academic specialised text (see [5], [6] on the differences/similarities: L1 vs L2 writing). Since a working definition had to be put forward, in the case of L1 English academic text samples we proposed a possible checklist to be considered. Thus, from the point of view of the text producer(s) the more criteria they comply with the better: affiliation with a university from a country whose sole official language is English, native speaker(s) or equivalent (if Bio/CV/language history available), journal impact factor. We have to deal here with the possibility of a multi-authored text, where the producers' L1 is different and also consider variation in the linguistic level [7], as in other cases, "proficiency levels appear to vary a great deal" [1].
As concerns English as L2, textual intervention might prove to be an issue. Since professional translation services are but a click away, it may prove a daunting task to determine the extent to which the English academic text belongs to its author(s)' voice. Moreover, the editorial process of established journals may prompt resorting to amendments to the original voice (suggestions for revision, multiple submissions, professional editing). For further research stages, solutions need to be sought and well thought to mitigate these issues.

4.2 Legal aspects

Another challenge we encountered refers to the legal aspects concerning the corpus data: copyright issues. Most of our linguistic sources are online journals that adopt an open-access policy. Many articles published in such journals are distributed under a Creative Commons license that grants copyright permissions and offers a standard set of terms and conditions that licensors may impose. In addition, we encountered another copyright issue regarding subscription-based journals, indexed in international databases (EBSCO, ERIH+, CEEOL etc.). They have all rights reserved, which means we have to obtain the copyright holder's permission to use their work in our corpus. Asking for individual permission might hamper progress in DACRE – given the communication workload – but, as such articles are valuable for our research, we are considering this approach as well.

4.3 Availability

In terms of availability of the resources, the Linguistics field seems to be privileged, while in domains such as IT or even Economics (Fig.2) there are almost no valuable journals that publish in Romanian, considering our main criteria: quality of the publications, indexation and publishing date. One possible solution was to extend the publication date period of the articles, looking for papers written up to 10 years earlier (with no result). Another solution was to search also for books published in Romanian by scholars (with some results, but difficult to download or access the source).

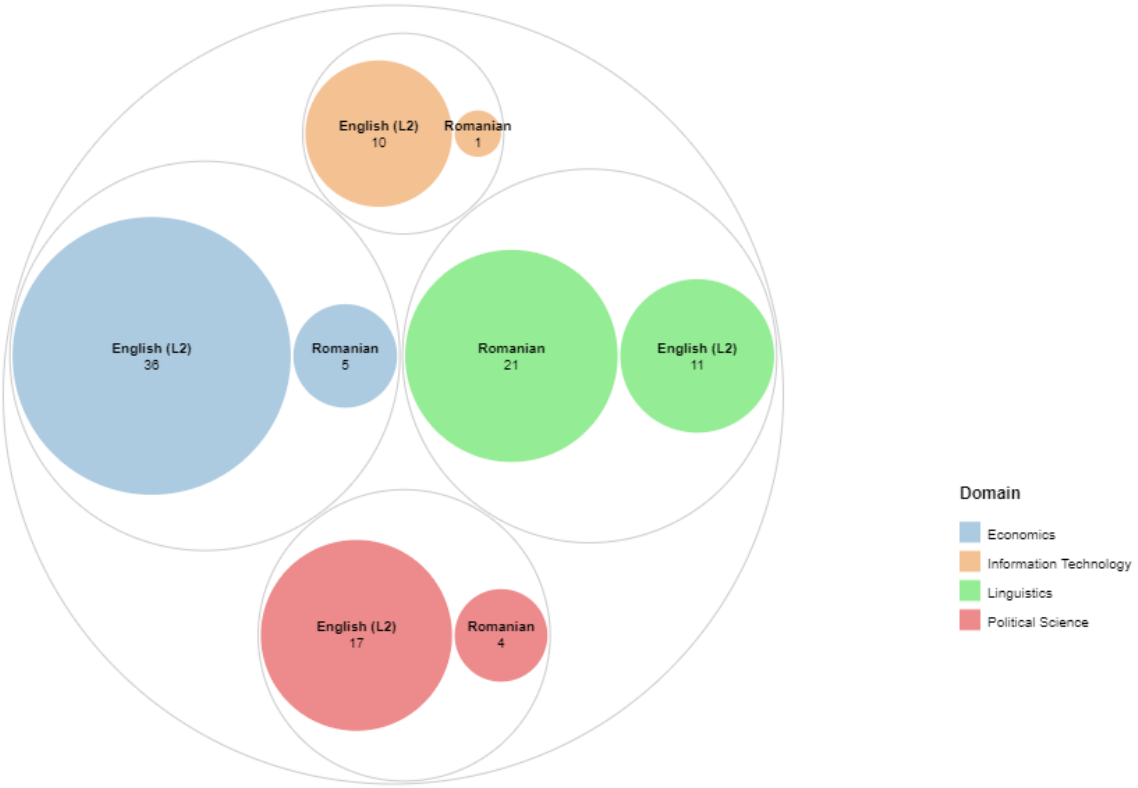

Fig.2: Status of scientific article collection in the DACRE corpus

4.4 Using automated models

The main challenge in the automatic collection of scientific articles from online libraries is that, even if some platforms allow manual downloading, they may not allow scraping/other computational methods to retrieve documents. Web scraping can be detected from online behaviour (e.g. repetitive patterns, multiple page visits in a short period of time) primarily by machine learning algorithms [8].

In the process of automating the extraction of articles from full journal volumes and issues containing multiple works, processing the documents despite their different formats is challenging. In order to detect each article bound from the pdf documents, the page numbers for each paper are extracted through computational models from the table of contents. It is a tedious task, as the table of contents varies between journals and it is hard to develop an algorithm capable of processing all the different formats of the volumes.

4.5 Multi-authorship

Another challenge regarding articles written in English L2 by Romanian researchers has been the tendency towards multi-authorship, which has been encountered mostly in the IT field: a significant amount of IT articles have been written in collaboration with other (especially foreign) researchers. We have several hypotheses regarding potential causes for multi-authorship: Romanian IT researchers tend to publish articles via conferences, or the collaborations provide more access to international journals. As such, the issue of multi-authorship might affect our data collection process.

5. Principles of a model for expert corpus building and conclusions

By describing, comparing and analyzing data collection barriers, we can now propose a model for expert corpus building in English vs in low-resource languages such as Romanian (Table 2):

Data collection model for English L1 expert corpora	Data collection model for low-resource-language expert corpora
• (a) selection of specialised journals with open-access • (b) automated article extraction • (c) linguistic analysis on separate corpus batches and the whole corpus	• (a1) selection of specialised open-access journals // • (a2) selection of specialised journals without open-access • (b1) automated article extraction // • (b2) contact non-open-access resources administrators and manual extraction • (c) evaluation and categorisation in quality groups • (d) linguistic analysis on separate corpus batches and the whole corpus

Table 2: Data collection models

The preliminary corpus collection stages in DACRE revealed the prevalence of a vicious cycle that affects corpus-based research in low-resource languages: for example, the high degree of challenges in building expert corpora originates in the lack of peer-reviewed publications in the mother tongue and the difficulty to identify expert-level English L2 writings. Thus, projects such as DACRE are essential in providing methodologies and instruments for the academic and professional community.

Acknowledgement

DACRE
Discipline-specific expert academic writing in Romanian and English: corpus-based contrastive analysis models

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] Mauranen, A., Ranta, E. “English as an Academic lingua franca – the ELFA project”, *Nordic Journal of English Studies*, 7(3), 2008, 199-202.
[2] Bercuci, L., Chitez. M. “A corpus analysis of argumentative structures in ESP writing”, *International Online Journal of Education and Teaching (IOJET)*, 6(4), 2019, 733-747.
[3] Băniceru, C., Tucan, D. “Perceptions About “Good Writing” and “Writing Competences” in Romanian Academic Writing Practices: A Questionnaire Study”. In Chitez, M., Doroholschi, C.I., Kruse, O., Salski, Ł., Tucan, D. (Eds.), *University Writing in Central and Eastern Europe: Tradition, Transition, and Innovation*, Cham, Springer, 2018, 103-112.
[4] O’Sullivan, Í. “Using corpora to enhance learners' academic writing skills in French”, *Revue française de linguistique appliquée*, 2(2), 2010, 21-35.
[5] Silva, T. “Toward an Understanding of the Distinct Nature of L2 Writing: The ESL Research and Its Implications”, *TESOL Quarterly*, 27(4), 1993, 657-677.
[6] Silva, T. “Differences in ESL and native-English speaker writing: The research and its Implications”. In Severino, C., Guerra, J.C., Butler, J.E. (Eds.), *Writing in multicultural settings*, New York, Modern Language Association of America, 1997, 209-219.

[7] Yilmaz, S., Römer, U. “A corpus-based exploration of constructions in written academic English as a lingua franca”. In Römer, U., Cortes, V., Friginal, E. (Eds.), *Advances in Corpus-based Research on Academic Writing. Effects of discipline, register, and writer expertise*, Amsterdam/Philadelphia, John Benjamins, 2020, 59-88.
[8] Meschenmoser, P., Meuschke, N., Hotz, M., Gipp, B. “Scraping scientific web repositories: Challenges and solutions for automated content extraction”, *D-Lib Magazine*, 22(9/10), 2016.