Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4739–4746
Marseille, 20-25 June 2022

EXPRES Corpus for A Field-specific Automated Exploratory Study of L2 English Expert Scientific Writing

Ana-Maria Bucur[1,2], **Madalina Chitez**[2], **Valentina Muresan**[2], **Andreea Dinca**[2], **Roxana Rogobete**[2]
[1] University of Bucharest [2] West University of Timișoara, Romania
ana-maria.bucur@drd.unibuc.ro
{madalina.chitez, valentina.muresan, andreea.dinca, roxana.rogobete}@e-uvt.ro

Abstract

Field Specific Expert Scientific Writing in English as a *Lingua Franca* is essential for the effective research networking and dissemination worldwide. Extracting the linguistic profile of the research articles written in L2 English can help young researchers and expert scholars in various disciplines adapt to the scientific writing norms of their communities of practice. In this exploratory study, we present and test an automated linguistic assessment model that includes features relevant for the cross-disciplinary second language framework: Text Complexity Analysis features, such as Syntactic and Lexical Complexity, and Field Specific Academic Word Lists. We analyse how these features vary across four disciplinary fields (Economics, IT, Linguistics and Political Science) in a corpus of L2-English Expert Scientific Writing, part of the EXPRES corpus (Corpus of Expert Writing in Romanian and English). The variation in field specific writing is also analysed in groups of linguistic features extracted from the higher visibility (Hv) versus lower visibility (Lv) journals. After applying lexical sophistication, lexical variation and syntactic complexity formulae, significant differences between disciplines were identified, mainly that research articles from Lv journals have higher lexical complexity, but lower syntactic complexity than articles from Hv journals; while academic vocabulary proved to have discipline specific variation.

Keywords: Text Complexity Analysis, Academic Vocabulary, Expert Scientific Writing in English, EXPRES Corpus, Indexed Journal Writing

1. Introduction

Developing proficient writing skills in English has been a debated subject over the last decades, since English has become the ”the main *lingua franca* for research networking and scientific communication” (Pérez-Llantada, 2012). Researchers and professionals in the disciplines are often hampered in their endeavours to disseminate scientific research results because of insufficient academic writing skills, and developing them is a challenge regardless of the field of interest. However, writing in the disciplines, while using a foreign language, strongly relies on an understanding of the writing practices of each particular field (Bazerman, 1991).

Research articles (RAs), as an academic genre, hold a central place in academia, as they are the main form of scientific communication (Swales, 1990). Although it has a similar function across disciplines, namely that of communicating research findings, the research article differs substantially from one discipline to another. The current investigation seeks to understand if there are significant linguistic differences between several field-specific discourse communities (Swales, 1990; Hyland, 2008), such as Linguistics, Economics, Information Technology, and Political Sciences, by analysing specific research articles written by a particular group of English-L2 scholars (i.e. Romanian scholars). Being an exploratory study, it aims to compare the linguistic profiles and patterns within expert academic writing published in two different categories of Romanian scientific journals, according to their international indexing: higher visibility journals (ISI/Web of Science, EBSCO, SCOPUS, ERIHPLUS) and journals present in less prestigious IDBs, search engines and citation databases (assumed to have lower visibility). The analysed collection of academic papers included in this paper is part of EXPRES (Corpus of Expert Writing in Romanian and English). The EXPRES expert writing corpus, the first such corpus reflecting the field specific academic writing profile of Romanian researchers, contains peer-reviewed research articles written between 2017 and 2021 in the aforementioned disciplines (Linguistics, IT, Political Sciences and Economics). The comparison of the two categories of journals is performed using lexical and syntactic complexity metrics and an analysis of predominant academic vocabulary.

If other studies examine differences between native and non-native English-speaking scholars (Lu et al., 2019), this paper analyses, in the first place, the differences in the linguistic profile of the field-specific expert academic writing of a particular group of non-native researchers (i.e. Romanians), and, secondly the distribution of such differences depending on the journal's international indexing.

Several main research questions will be taken into consideration:

- Are there significant field specific differences, in the case of the selected four fields (Economics, IT, Linguistics and Political Science), regarding the lexical and syntactic complexity of the L2 English expert scientific writing?

- Are the features mentioned in (a) distributed differently in the higher visibility (Hv) vs lower visibility (Lv) journals?
- Are there any identifiable field-specific academic words characterizing the expert academic writing of a particular set of L2 English users (e.g. Romanian scholars)?
- What do automated complexity analysis tools tell us about expert writing in the above mentioned four fields and, more specifically, about how English is used for research article writing?

In order to answer these, we will scrutinize the linguistic complexity of the collected data. Through this contribution, we are interested in evaluating the linguistic profile of the academic expert writing in English as a lingua franca specific to a particular L1 scientific writing community (i.e. Romanian scholars) in order to be able to validate an automated assessment model that can be used for other L1 scientific writing groups.

2. Related Work

The aim of our study is to analyse the results of an automated linguistic assessment study conducted on a corpus of expert writing in English L2, i.e. scientific articles, in different disciplines. Our specific focus is on the result correlation with the international indexing level of the article publication source. Expert scientific writing is referred to as "articles from peer-reviewed, top-rated journals", as exemplified in Larsson's study on the LOCRA corpus[1] (Louvain Corpus of Research Articles) (Larsson, 2016). Broadly speaking, expert scientific writing encompasses different genres labelled as "published scientific writing" (Salazar, 2014), although scientific or research articles are the preferred text types that fall under this academic writing subgroup. Extending the definition, expert corpora contain "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" (Rogobete et al., 2021).

Automatic evaluation methods of writing in English L2 or any other second / foreign language have been developed, tested and analysed in numerous studies which have approached "linguistic complexity as a multilevel phenomenon" (Green, 2019). The assessment formulae include, predominantly, multiple complexity features (Okinina et al., 2020; Housen et al., 2019), syntactic complexity and sophistication markers (Kyle and Crossley, 2017) or lexical complexity, richness or density (Lu, 2014). The decision to include particular features depends on the typology of data and research questions. Thus, replications or variations of the multidimensional analysis (MDA) (Biber, 1992), "which reduces large sets of linguistic variables, typically around 150 to meaningful dimensions of correlated variables" (Green, 2019), have been rather relevant for register variations. In order to compare writing in the disciplines, discriminant function analysis (DFA) seems to be the alternative (Egbert and Biber, 2018), with its "emphasis on highlighting differences, is conceptually well-aligned with disciplinary literacy" (Green, 2019). However, in order to conduct DFA based analyses, large amounts of metadata should be collected (e.g. grades, genres), which makes it quite difficult to employ for exploratory studies. Such studies have simplified the analysis model to include lexical complexity measures (e.g. lexical diversity and lexical density) and syntactic complexity measures (e.g. sentence length, ratio of subordination) (Khany and Kafshgar, 2016). In addition, academic words appear to be an important indicator of the discipline-specific linguistic profile of academic texts (Hyland and Tse, 2007). The procedure can be complemented by automated term extraction for disciplines (Periñán-Pascual, 2018).

Since intensive research involves knowledge transfer and dissemination, research production must be related to the audience (either experienced or novice) in order to contribute to better access of practitioners to novel inquiries. In the case of academic journals, multiple ranking metrics are used to assess quality, impact, and visibility as main recognition factors. Scholars aim to access highly ranked publications in order to acquire a wider visibility and increase their individual metrics that capture productivity, citation impact, and research output overall. However, a comparative analysis that focuses on journals with lower vs higher visibility leads to the following observation: even though a minor part of researchers tend to write their RAs in a language "as scientific as possible", difficult to understand (Gazni, 2011), potential wide readership can be obtained through an appropriate level for the general population, because there is no need to have an elitist perspective, but to share knowledge to the entire community. A study by Moohebat et al. (2015) demonstrates how lexical usage analyses can be used to train text classification models to distinguish between scientific writing in ISI versus non-ISI journals.

Our automated assessment model was tested for the EXPRES corpus, representing English-L2 expert writing produced by Romanian scholars. The text assessment measures (syntactic and lexical complexity, the use of words from the Academic Word List) were adapted from similar assessment models considered relevant for analyses within a cross-disciplinary second language framework.

3. Data

EXPRES Corpus The collection compiled for this study is part of EXPRES (*Corpus of Expert Writing in Romanian and English*), a discipline-specific academic writing corpus consisting of research articles in peer-reviewed journals, aiming to support Romanian fac-

[1] https://uclouvain.be/en/research-institutes/ilc/cecl/locra.html

	#Articles		#Words/Article		#Unique words/Article	
Domain	**Hv**	**Lv**	**Hv**	**Lv**	**Hv**	**Lv**
Economics	465	157	3675 (1559)	3038 (1156)	901 (311)	824 (246)
IT	269	58	3788 (1160)	2259 (1110)	895 (239)	598 (258)
Linguistics	143	350	5552 (2212)	3528 (1369)	1470 (476)	1256 (431)
Political Science	118	21	4801 (1881)	5692 (3095)	1314 (421)	1399 (552)

Table 1: Summary of the corpus. Hv - higher visibility journals. Lv - lower visibility journals. We report the mean and standard deviation for the number of words and number of unique words in articles.

ulty members, professionals, researchers and students in order to communicate their research findings. EXPRES has a manifold corpora typology, focusing on four fields of academic research (Linguistics, Political Sciences, Economics and Information Technology) and two languages, but with different user levels: English L1 (articles written by native-like experts, published in peer-reviewed journals in English-speaking countries), English L2 (articles written by Romanian experts using English as a Foreign Language), Romanian L1 (articles written by Romanian experts in their mother tongue).

Regarding the research articles written in Romanian, we have to mention that there are only a few Romanian-language publications with high visibility (Rogobete et al., 2021).

Since English is widely used within Romanian higher education institutions – both as an instruction medium and for research publication purposes, the English L2 sub-corpus seemed to better adapt to automated extraction models of online Romanian journal articles.

The English L2 articles were selected from two categories of Romanian scientific journals, according to their international indexing: higher visibility journals (ISI/Web of Science, EBSCO, SCOPUS, ERIHPLUS) and lower visibility journals, present in fewer or less prestigious IDBs, search engines and citation databases (assumed to have a lower impact). All of them are peer-reviewed research articles written between 2017 and 2021, specific to the aforementioned domains.

Collecting the data The academic articles were collected manually and automatically by downloading the PDFs of the papers from the journals' websites. The automated method consisted of scraping the URLs of the articles using the Python crawling framework *scrapy*[2]. The process of gathering the URLs was challenging as the journals' websites have very different layouts. The most challenging features of the websites were: using pictures with text instead of the actual text, using HTML frames and hosting the articles on external platforms, some of which were unavailable. The PDFs of the papers were downloaded from the URLs. Using the Java library *Cermine*[3], the contents of the PDFs were extracted. We filtered the data of publishing in order to obtain our subset of articles published since 2017 and in which all the authors have Romanian names (filtered using the list of Romanian names from Wikipedia[4]).

A summary of the sub-corpus used in our analyses is presented in Table 1. It consists of 995 articles from higher visibility (Hv) journals and 586 from journals with lower visibility (Lv).

For data selection, a number of criteria were used, including the author's identity (namely Romanian authors to ensure the appurtenance to an English as L2 community) and expertise (academics), the availability and status of expert writing samples (opting for open source journals, operating under the Creative Commons license) and the journal impact factor.

As seen in Table 1, the articles have various lengths. The variation in the number of words in an article may be an effect of the word limit imposed by some publishers but not others. The articles published in Hv journals have a higher word count than articles from Lv journals for Economics, IT and Linguistics. For the field of Political Science, there is no direct correspondence between paper length and journal visibility, for example, an Lv journal requires longer paper submissions than any of the Hv journals in our corpus. As regards the field of Linguistics, there is a great variety of word limits among Hv journals, ranging from about 4000 words to about 8000 words, while this limit may be even higher in some journals, 12.000 or even 16.000 words.

Additionally, we have to mention that in the case of Economics and Information Technology, the formulas are not included in the word count.

4. Methods

To understand how L2 English is used by a specific group of academics (i.e. Romanian scholars) in research paper writing, different automatic measures are explored in this work. In this section, we describe the methods used for analysing the lexical and syntactic complexity as well as the academic vocabulary used in the research papers from our sub-corpus. Welch's t-test was used for measuring the statistical significance of the differences between the metrics from Hv and Lv journals from the four fields.

4.1. Lexical Complexity Analysis

For analysing the lexical complexity of research articles, we computed several measures for lexical density, sophistication and variation.

[2] https://scrapy.org/

[3] https://github.com/CeON/CERMINE

[4] https://ro.wikipedia.org/wiki/Listă_de_nume_românești

Lexical density is computed as the ratio of lexical words to all the words in a document. Lexical words are defined as "nouns, adjectives, verbs (excluding modal verbs, auxiliary verbs, *be* and *have*), and adverbs with an adjectival base, including those that can function as both an adjective and adverb (e.g., *fast*) and those formed by attaching the *–ly* suffix to an adjectival root (e.g., *particularly*)" (Lu, 2012).

Lexical sophistication is computed as the ratio of sophisticated words from all the words in the document. The sophisticated words are words not appearing on the list of 2,000 most frequent words in the British National Corpus (Leech et al., 2014). While in learner corpora this sophistication is rather rare (Read, 2000), it is expected that in expert corpora (such as EXPRES) the proportion should be higher.

Lexical variation assesses the diversity of the words used in a document. The most common evaluated indices are textual lexical diversity (MTLD), vocabulary diversity (Vocd-D), Uber Index (Uber), and squared verb variation (SVV) (Kalantari and Gholami, 2017).

The lexical complexity measures were computed using the Lexical Complexity Analyser (LCA)[5] (Lu, 2012).

4.2. Syntactic Complexity Analysis

Trying to identify the characteristics of expert academic writing as L2 production, we also performed a syntactic complexity analysis and computed measures for the length of the production unit, the amount of subordination, the amount of coordination, the degree of phrasal complexity and the overall sentence complexity using the L2 Syntactical Complexity Analyser (L2SCA)[6] (Lu, 2010). The L2SCA tool uses the Stanford parser (Klein et al., 2003) for parsing the documents and identifying the production units. Tregex (Levy and Andrew, 2006) is used for counting the different production units.

4.3. Academic Vocabulary

For identifying the lexical preferences within the academic writing genre we used the Academic Word List (AWL)[7] (Coxhead, 2000) to compute the percentage of academic words occurring in the articles across the four disciplines. AWL contains 570 word families, but it does not include words from the list of 2,000 most frequent words in English. Furthermore, it was compiled from a corpus of 28 subject areas, thus AWL relies on vocabulary which covers all the fields in the EXPRES corpus, most in direct correspondence - Linguistics, Economics, Computer Science, while Political Science is only being indirectly covered by areas such as Rights and Remedies, Constitutional Law and Sociology.

5. Results and Discussion

The results from analysing the lexical and syntactic complexity and computing the percentage of academic words are presented below. We compare articles from Hv and Lv journals and show the differences between the four domains: Economics, IT, Linguistics, and Political Science.

5.1. Lexical Complexity Analysis

Metric	Domain	Mean (SD)		*p*-value
		Hv	**Lv**	
Lexical	**Economics**	0.57 (0.03)	0.58 (0.04)	$<$ **0.001**
density	**IT**	0.60 (0.04)	0.60 (0.04)	0.993
N_{lex}/N	**Linguistics**	0.58 (0.04)	0.69 (0.14)	$<$ **0.001**
	Political Science	0.62 (0.10)	0.59 (0.08)	0.280
Lexical	**Economics**	0.34 (0.06)	0.35 (0.06)	0.116
sophistication	**IT**	0.40 (0.07)	0.41 (0.08)	0.534
N_{slex}/N_{lex}	**Linguistics**	0.50 (0.08)	0.65 (0.22)	$<$ **0.001**
	Political Science	0.48 (0.18)	0.44 (0.17)	0.334
Lexical sophistication-II	**Economics**	0.40 (0.06)	0.40 (0.06)	0.498
	IT	0.44 (0.05)	0.43 (0.07)	0.103
T_s/T	**Linguistics**	0.56 (0.08)	0.68 (0.21)	$<$ **0.001**
	Political Science	0.55 (0.17)	0.52 (0.15)	0.325
Verb	**Economics**	0.14 (0.04)	0.14 (0.03)	0.076
sophistication	**IT**	0.13 (0.03)	0.14 (0.04)	**0.044**
T_{sverb}/N_{verb}	**Linguistics**	0.20 (0.05)	0.39 (0.23)	$<$ **0.001**
	Political Science	0.23 (0.10)	0.21 (0.15)	0.710
Corrected VS1	**Economics**	1.75 (0.57)	1.65 (0.46)	**0.028**
	IT	1.75 (0.41)	1.40 (0.42)	$<$ **0.001**
$T_{sverb}/\sqrt{2N_{verb}}$	**Linguistics**	3.20 (0.85)	3.85 (1.53)	$<$ **0.001**
	Political Science	3.14 (1.18)	2.98 (1.20)	0.552
Verb sophistication-II	**Economics**	6.74 (4.77)	5.85 (3.31)	**0.009**
	IT	6.48 (3.05)	4.26 (2.53)	$<$ **0.001**
T^2_{sverb}/N_{verb}	**Linguistics**	21.88 (11.02)	34.37 (25.67)	$<$ **0.001**
	Political Science	22.57 (17.78)	20.46 (20.23)	0.658

Table 2: Lexical density and sophistication. T = #word types, N = #word tokens, s = sophisticated words, lex = lexical words, $slex$ sophisticated lexical words, $sverb$ = sophisticated verbs.

Metric	Domain	Mean (SD)		*p*-value
		Hv	**Lv**	
Lexical word	**Economics**	0.36 (0.08)	0.39 (0.07)	$<$ **0.001**
variation	**IT**	0.33 (0.06)	0.38 (0.10)	$<$ **0.001**
T_{lex}/N_{lex}	**Linguistics**	0.39 (0.09)	0.46 (0.08)	$<$ **0.001**
	Political Science	0.38 (0.07)	0.36 (0.07)	0.130
Verb variation	**Economics**	0.48 (0.10)	0.51 (0.08)	$<$ **0.001**
T_{verb}/N_{verb}	**IT**	0.41 (0.07)	0.47 (0.10)	$<$ **0.001**
	Linguistics	0.48 (0.10)	0.63 (0.15)	$<$ **0.001**
	Political Science	0.51 (0.08)	0.49 (0.12)	0.493
Noun variation	**Economics**	0.34 (0.08)	0.37 (0.07)	$<$ **0.001**
	IT	0.31 (0.07)	0.36 (0.10)	$<$ **0.001**
T_{noun}/N_{lex}	**Linguistics**	0.39 (0.09)	0.46 (0.09)	$<$ **0.001**
	Political Science	0.38 (0.07)	0.35 (0.07)	0.068

Table 3: Lexical variation. T = #word types, N = #word tokens, lex = lexical words.

Tables 2 and 3 display the results obtained from the Lexical Complexity Analysis, with two major trends emerging from the data. Interestingly, and rather surprisingly, the lower visibility journals from Economics, IT and Linguistics show significantly higher scores for all the lexical complexity metrics investigated, when compared to the higher visibility journals, suggesting that Lv articles are more complex than the Hv articles. The most statistically significant difference is noticed in the Lv Linguistics corpus, which scores considerably higher than all the other corpora, regardless of the journals' visibility. In contrast, in the case of Political Sciences, the Hv corpus shows higher values in comparison with the Lv corpus.

[5] http://www.personal.psu.edu/xxl13/downloads/lca.html

[6] https://aihaiyang.com/software/l2sca/

[7] https://www.wgtn.ac.nz/lals/resources/academicwordlist

Metric	**Domain**	**Mean (SD)**		p-value
		Hv	**Lv**	
Length of production unit				
Mean length of sentence	**Economics**	29.58 (5.91)	27.95 (6.20)	**0.004**
	IT	24.81 (4.29)	24.98 (6.20)	0.831
	Linguistics	27.26 (5.52)	26.89 (7.84)	0.552
	Political Science	28.72 (6.28)	27.28 (6.64)	0.366
Mean length of clause	**Economics**	16.40 (2.93)	16.84 (2.75)	0.087
	IT	14.70 (1.98)	14.30 (2.16)	0.203
	Linguistics	14.16 (2.79)	15.96 (3.63)	< **0.001**
	Political Science	16.27 (3.00)	14.95 (2.96)	0.071
Mean length of T-unit	**Economics**	27.13 (5.03)	26.72 (5.88)	0.432
	IT	22.92 (3.64)	22.67 (4.91)	0.725
	Linguistics	25.61 (4.96)	26.93 (6.55)	**0.015**
	Political Science	29.00 (6.31)	25.83 (6.39)	**0.044**
Amount of subordination				
Number of clauses per T-unit	**Economics**	1.67 (0.23)	1.60 (0.33)	**0.018**
	IT	1.56 (0.18)	1.59 (0.25)	0.506
	Linguistics	1.83 (0.26)	1.71 (0.27)	< **0.001**
	Political Science	1.79 (0.26)	1.73 (0.23)	0.259
Complex T-unit ratio	**Economics**	0.43 (0.10)	0.39 (0.10)	< **0.001**
	IT	0.39 (0.09)	0.37 (0.12)	0.267
	Linguistics	0.47 (0.10)	0.38 (0.15)	< **0.001**
	Political Science	0.45 (0.12)	0.47 (0.12)	0.540
Number of dependent clauses per clause	**Economics**	0.37 (0.07)	0.34 (0.08)	< **0.001**
	IT	0.33 (0.07)	0.32 (0.09)	0.254
	Linguistics	0.40 (0.07)	0.32 (0.11)	< **0.001**
	Political Science	0.38 (0.09)	0.38 (0.08)	0.697
Number of dependent clauses per T-unit	**Economics**	0.63 (0.20)	0.56 (0.23)	**0.002**
	IT	0.53 (0.17)	0.52 (0.21)	0.820
	Linguistics	0.75 (0.23)	0.58 (0.28)	< **0.001**
	Political Science	0.71 (0.24)	0.66 (0.23)	0.454
Amount of coordination				
Number of coordinate phrases per clause	**Economics**	0.53 (0.19)	0.58 (0.21)	**0.002**
	IT	0.38 (0.15)	0.38 (0.18)	0.888
	Linguistics	0.37 (0.16)	0.26 (0.21)	< **0.001**
	Political Science	0.44 (0.21)	0.36 (0.16)	0.065
Number of coordinate phrases per T-unit	**Economics**	0.86 (0.29)	0.92 (0.33)	0.060
	IT	0.59 (0.23)	0.59 (0.26)	0.947
	Linguistics	0.67 (0.27)	0.48 (0.40)	< **0.001**
	Political Science	0.77 (0.37)	0.62 (0.26)	**0.028**
Number of T-units per sentence	**Economics**	1.09 (0.09)	1.05 (0.15)	**0.002**
	IT	1.08 (0.09)	1.10 (0.10)	0.262
	Linguistics	1.07 (0.12)	1.00 (0.15)	< **0.001**
	Political Science	1.00 (0.12)	1.00 (0.10)	**0.011**
Degree of phrasal complexity				
Number of complex nominals per clause	**Economics**	2.41 (0.53)	2.48 (0.50)	0.099
	IT	1.99 (0.36)	1.90 (0.39)	0.093
	Linguistics	2.02 (0.47)	1.93 (0.44)	0.051
	Political Science	2.22 (0.49)	2.08 (0.45)	0.223
Number of complex nominals per T-unit	**Economics**	3.99 (0.92)	3.97 (1.18)	0.850
	IT	3.11 (0.65)	3.01 (0.81)	0.411
	Linguistics	3.66 (0.84)	3.32 (1.06)	< **0.001**
	Political Science	3.96 (1.00)	3.63 (1.12)	0.220
Number of verb phrases per T-unit	**Economics**	2.31 (0.36)	2.24 (0.45)	0.060
	IT	2.22 (0.33)	2.19 (0.45)	0.617
	Linguistics	2.39 (0.42)	2.17 (0.53)	< **0.001**
	Political Science	2.40 (0.44)	2.33 (0.38)	0.446
Overall sentence complexity				
Number of clauses per sentence	**Economics**	1.82 (0.31)	1.68 (0.36)	< **0.001**
	IT	1.70 (0.26)	1.76 (0.38)	0.255
	Linguistics	1.96 (0.39)	1.73 (0.47)	< **0.001**
	Political Science	1.79 (0.35)	1.84 (0.30)	0.506

Table 4: Syntactic complexity measures.

Lexical density and sophistication. The lexical density metric does not show notable differences between Economics, IT, Political Science Lv and Hv articles. What stands out in this category is the Lv Linguistics' significantly higher score as compared with Hv Linguistics. Turning to the next two metrics, the Linguistics and the Political Science corpora display a higher degree of lexical and verb sophistication when compared to Economics and IT, regardless of the journals' visibility.

Lexical variation. In line with the general trend, the scores of the Lv Economics, IT and Linguistics corpora are slightly higher than the Hv corresponding corpora. In addition, what stands out in Table 3, however, is that the lower visibility Linguistics corpus shows significantly higher scores for each metric investigated as compared with the other corpora. Secondly, the IT higher visibility corpus scores lowest in all categories,

Word (number of occurrences)			
Economics	**IT**	**Linguistics**	**Political Science**
economic (6661)	data (4159)	text (1321)	economic (960)
financial (4277)	process (1445)	cultural (1210)	policy (671)
research (3711)	network (1116)	culture (1072)	security (580)
data (3673)	project (1077)	analysis (866)	cultural (505)
analysis (3216)	research (936)	research (749)	process (489)
economy (2884)	security (916)	process (742)	economy (377)
impact (2360)	technology (847)	context (735)	culture (369)
period (2219)	devices (807)	structure (694)	global (366)
process (2133)	approach (786)	texts (673)	energy (365)
environment (2081)	analysis (732)	found (670)	role (362)
resources (1995)	method (725)	focus (632)	approach (362)
significant (1770)	image (643)	specific (625)	community (356)
factors (1690)	methods (635)	author (614)	research (352)
variables (1676)	design (633)	perspective (613)	analysis (336)
sustainable (1658)	components (587)	role (611)	context (300)

Table 5: Top 15 words from AWL in the four domains: Economics, IT, Linguistics, Political Science

a tendency that is maintained by both Lv and Hv IT corpora in the Syntactic Complexity Analysis as well.

5.2. Syntactic Complexity Analysis

The results of the Syntactic Complexity Analysis are presented in Table 4. As illustrated in this table syntactic complexity is investigated as a multifaceted construct considering a larger number of metrics, including phrasal complexity measures in line with newer research claims (Biber et al., 2011).

It must be noticed that in spite of the obvious differences in levels of syntactic complexity between the expert writing samples belonging to the four fields, the metrics indicate their appurtenance to the category of proficient writing in L2, with the writing in fields of Economics and Political Science, in Hv journals, displaying the highest degree of syntactic complexity for most of the measures considered. In the case of Hv journal articles in the field of Linguistics, there can be observed that these have a higher overall sentence complexity than those in IT and Political Science or Economics, their syntactic complexity following the more recent preference for a higher level of embedding and higher phrasal-level complexification (Norris and Ortega, 2009).

Comparing the syntactic complexity features across both the Hv journals and the Lv ones, it can be stated that the articles from the IT field have the lowest syntactic complexity (except for the number of coordinate phrases per clause and the number of T-units per sentence). This finding could be explained by taking into account the discursive features of writing in this field (text acting as a support for the formulas/ equations/ code etc.). In the same comparison between Hv and Lv journal articles, in the case of Linguistics, even if the articles from Lv journals display a greater degree of lexical density and sophistication, they have lower syntactic complexity in contrast with those in Hv journals (except for the mean length of clause).

5.3. Academic Vocabulary

Domain	Mean (SD)		*p*-value
	Hv	Lv	
Economics	13.70% (2.82%)	13.92% (2.98%)	0.425
IT	**13.04% (2.67%)**	**11.53% (3.65%)**	**0.004**
Linguistics	7.96% (2.74%)	7.84% (3.65%)	0.705
Political Science	10.90% (3.01%)	11.43% (3.31%)	0.504

Table 6: Occurrence of words from the Academic Word List across disciplines.

In Table 5 and 6 we present the results of our analysis concerning the occurrence of words from the Academic Word List in our sub-corpus across the four fields: Economics, IT, Linguistics, Political Science.

Table 6 reveals that in the articles from the field of Linguistics, there seems to be a limited occurrence of words from the AWL vocabulary in comparison with the estimated mean of 8.5% to 10% in academic texts (Coxhead and Nation, 2001), a tentative explanation being the greater linguistic flexibility of experts in this field, who are less likely to opt for prescriptive, formulaic language. Additionally, an explicit Introduction, Methods, Results, and Discussion (IMRaD) structure for research papers may be less frequent in articles belonging to the sub-fields of literary and cultural studies (included in the broader field of Linguistics in our corpus).

Similarly, in her study on the genre- and discipline-specific recurrent expressions (lexical bundles), Dontcheva-Navratilova (2012) discusses a case of lower frequency of formulaic language in articles from the sub-field of literature or cultural studies in comparison with expert writing in the field of language studies (Linguistics).

Table 5 reveals that, regardless of the field, there are few common individual lexical items used in academic texts (such as ’research’, ’analysis’) which behave similarly across disciplines.

6. Conclusion

The present work mapped the characteristics of scientific papers written by Romanian academics from the EXPRES corpus compiled within the DACRE project. A total of 995 articles from higher visibility journals and 586 from lower visibility journals were analysed. The linguistic profile of expert writing in Hv and Lv journals was a result of comparisons relying on lexical and syntactic complexity levels, and on measuring the occurence of words from the Academic Word List. The findings presented in this work reveal great differences between the research papers from the four fields (Economics, IT, Linguistics and Political Science) and between the two categories of Romanian scientific journals, according to their international indexing (higher visibility journals and lower visibility journals).

In sum, for the field of **Linguistics**, there are greater differences between articles from Hv and Lv journals, articles from the latter category being more lexically sophisticated, thus harder to read and understand. However, this is compensated by the lower syntactic complexity levels of articles in LV journals in comparison with the articles from Hv journals (except for the mean length of clause).

The articles from the field of **Political Sciences** display higher lexical complexity compared to articles from the fields of Economics and IT. For this domain, there are no significant differences between articles across the indexed sources.

Articles in the field of **Economics** are characterised by lower lexical variation than those in Linguistics and Political Sciences but higher than those in IT.

Furthermore, the articles in the field of **IT** have the lowest lexical density and sophistication. We may conclude that the papers in the field of IT are the easiest to read and comprehend as they present the lowest lexical and syntactic complexity levels, thus making the textual items understandable by a wider audience; however, this apparent simplification is compensated by the extra-textual items (formulas/ equations/ coding/ symbols), not computed here.

The novelty of the present study consists, firstly, in compiling and analyzing the EXPRES corpus focused on expert writing, instead of learner corpora, previously used in extensive research studies on academic writing [8]. Secondly, the investigation aimed at comparing writing in four different fields in terms of lexical and syntactic complexity. Since "L2 writing quality [...] is a function of both writing ability and language proficiency" (Yang et al., 2015), the relation between the effect of a field of research, on the one hand, and linguistic complexity, writing performance and quality, on the other, is less argued, our contribution aimed to fill the research gap in comparing Hv/Lv journals in English L2. Additionally, our research aimed at drawing attention upon the need for researchers in particular fields to make adjustments to international writing norms in order to make their findings and results more visible. Finally, our findings may indicate a possible influence of L1 writing style as affecting L2 writing, even at expert level.

Acknowledgements

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).

7. Bibliographical References

Bazerman, C. (1991). The second stage in writing across the curriculum. *College English*, 53(2):209.

Biber, D., Gray, B., and Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in l2 writing development? *TESOL Quarterly*, 45(1):5–35.

Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, 15(2):133–163.

Coxhead, A. and Nation, P. (2001). The specialised vocabulary of english for academic purposes. *Research perspectives on English for academic purposes*, pages 252–267.

Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, 34(2):213–238.

Dontcheva-Navratilova, O. (2012). Lexical bundles in academic texts by non-native speakers. *Brno Studies in English*, (2).

Egbert, J. and Biber, D. (2018). Do all roads lead to rome?: Modeling register variation with factor analysis and discriminant analysis. *Corpus Linguistics and Linguistic Theory*, 14(2):233–273.

Gazni, A. (2011). Are the abstracts of high impact articles more readable? investigating the evidence from top research institutions in the world. *Journal of Information Science*, 37(3):273–281.

Green, C. (2019). A multilevel description of textbook linguistic complexity across disciplines: Leveraging nlp to support disciplinary literacy. *Linguistics and Education*, 53:100748.

Housen, A., De Clercq, B., Kuiken, F., and Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second language research*, 35(1):3–21.

Hyland, K. and Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, 41(2):235–253.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for specific purposes*, 27(1):4–21.

[8] the main results and the EXPRES corpus will be available on the DACRE platform with a complex search interface, to be launched at the end of 2022: `https://dacre.projects.uvt.ro/?lang=en`

Kalantari, R. and Gholami, J. (2017). Lexical complexity development from dynamic systems theory perspective: Lexical density, diversity, and sophistication. *International Journal of Instruction*, 10:1–18.

Khany, R. and Kafshgar, N. B. (2016). Analysing texts through their linguistic properties: A cross-disciplinary study. *Journal of Quantitative Linguistics*, 23(3):278–294.

Klein, D., Manning, C. D., et al. (2003). Fast exact inference with a factored model for natural language parsing. *Advances in neural information processing systems*, pages 3–10.

Kyle, K. and Crossley, S. (2017). Assessing syntactic sophistication in l2 writing: A usage-based approach. *Language Testing*, 34(4):513–535.

Larsson, T. (2016). The introductory it pattern: Variability explored in learner and expert writing. *Journal of English for Academic Purposes*, 22:64–79.

Leech, G., Rayson, P., et al. (2014). *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Levy, R. and Andrew, G. (2006). Tregex and tsurgeon: Tools for querying and manipulating tree data structures. In *LREC*, pages 2231–2234. Citeseer.

Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schnaars, M., and Zhang, C. (2019). Examining scientific writing styles from the perspective of linguistic complexity. *Journal of the Association for Information Science and Technology*, 70(5):462–475.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Lu, X. (2012). The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Lu, X. (2014). Lexical analysis. In *Computational Methods for Corpus Annotation and Analysis*, pages 67–93. Springer.

Moohebat, M., Raj, R. G., Kareem, S. B. A., and Thorleuchter, D. (2015). Identifying isi-indexed articles by their lexical usage: A text analysis approach. *Journal of the Association for Information Science and Technology*, 66(3):501–511.

Norris, J. M. and Ortega, L. (2009). Towards an organic approach to investigating caf in instructed sla: The case of complexity. *Applied linguistics*, 30(4):555–578.

Okinina, N., Frey, J.-C., and Weiss, Z. (2020). Ctap for italian: Integrating components for the analysis of italian into a multilingual linguistic complexity analysis tool. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7123–7131.

Pérez-Llantada, C. (2012). *Scientific discourse and the rhetoric of globalization: The impact of culture and language*. A&C Black.

Periñán-Pascual, C. (2018). Dexter: A workbench for automatic term extraction with specialized corpora. *Natural Language Engineering*, 24(2):163–198.

Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press.

Rogobete, R., Chitez, M., Mureșan, V., Damian, B., Duciuc, A., Gherasim, C., and Bucur, A.-M. (2021). Challenges in compiling expert corpora for academic writing support. In *Proceedings of the 11th International Conference The Future of Education (1-2 July 2021)*, pages 409–414. Filodiritto Editore.

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching*, volume 65. John Benjamins Publishing Company.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Yang, W., Lu, X., and Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28:53–67.