Disciplinary Variation in Syntactic Complexity: A Corpus Analysis of Professional Academic Writing

Javier Pérez-Guerra[a] and Elizaveta A. Smirnova[a,b]

[a] *University of Vigo, Campus Universitario, Vigo, E-36310, Spain*
[b] *HSE University, 38 Studencheskaya Street, Perm, 614070, Russia*

Abstract

This study deals with the analysis of syntactic complexity in professional academic writing and is based on a corpus of so-called 'hard' and 'soft' papers published in leading international journals. We aim at describing the main complexity features of academic discourse and testing the hypothesis that there is considerable disciplinary variation in linguistic complexity. We conclude that, first, clausal complexity strategies are more prevalent in the 'hard' sciences, while phrasal-complexity features dominate in the 'soft' ones. Second, the data reveal a continuum across subdisciplines within the broad categories of 'soft' and 'hard' genres with respect to the adoption of complexity strategies.

Keywords

Corpus analysis, disciplinary variation, academic discourse, academic writing, syntactic complexity

1. Introduction

The phenomenon of complexity has been extensively approached in corpus linguistics over the recent years. Specifically, the complexity of writing has been studied in terms of the comparison of L2 and L1 writing [e.g. 1], correlations between text complexity, language proficiency and task types [e.g. 2], and the development of text complexity after intensive instruction [e.g. 3]. However, complexity in professional academic writing has been relatively under-researched to date despite the potential pedagogical implications of such studies. In this respect, we contend that following the linguistic conventions of a particular discipline plays a crucial role in identifying the writers as experts in their own discourse communities [4]. From this perspective, a research article can serve as a benchmark for optimal academic writing, providing learners with "a rich and authentic introduction to the complexities and nuances of the genre" [5: 3]. This study reports the empirical analysis of linguistic complexity features which aims, first, to describe the complexity features of research articles written by professional authors and, second, to test the hypothesis that linguistic complexity varies across disciplines.

2. Data and methodology

The analysis of linguistic complexity in professional academic writing has been conducted on a 775,000-word corpus of research papers in four 'soft' arts and social sciences (business studies, linguistics, history and political science), and four 'hard' life and physical sciences (mathematics, engineering, chemistry and physics) which were published in leading peer-review journals indexed in Scopus Quartile 1, in 2016 and 2017. Once collected, the texts were manually cleared from tables, formulas, graphs, charts, metadata and reference lists for further analysis. The size and details of the corpus are given in Table 1.

IMS 2021 - International Conference "Internet and Modern Society", June 24-26, 2021, St. Petersburg, Russia
EMAIL: jperez@uvigo.es (A. 1); easmirnova@hse.ru (A. 2);
ORCID: 0000-0002-8882-667X (A. 1); 0000-0001-9307-6773 (A. 2)

Table 1
Corpus

Discipline	No. texts	Word totals	*Journals*
HARD SCIENCES			
Chemistry	16	97,947	*Cell Chemical Biology (CCB)*
			Chem
Physics	18	95,852	*Physics Letters B (PL)*
			Reviews in Physics (RP)
Mathematics	13	98,430	*Compositio Matematica (CM)*
			The Journal of Differential Geometry (JDG)
Engineering	17	99,003	*Automatica (Auto)*
			Materials Characterisation (MC)
Totals	64	391,232	
SOFT SCIENCES			
Business	10	95,350	*The Journal of Management (JM)*
			The Journal of Management Studies (JMS)
Linguistics	10	95,603	*Applied Linguistics (AL)*
			Lingua (Ling)
History	10	99,303	*Contemporary European History (CEH)*
			The Journal of Modern History (JMH)
Political science	11	93,366	*Political Analysis (PA)*
			World Politics (WP)
Totals	41	383,622	

In this study we undertake both the quantitative analysis of measures automatically generated by the complexity analyser and the qualitative scrutiny of a number of syntactic patterns associated with syntactic complexity. Firstly, to accomplish the quantitative analysis, the corpus texts were processed using Lu's L2 Syntactic Complexity Analyser (hereafter L2SCA). L2SCA provided the 14 indices given in Table 2 along with their descriptions, as in Lu [6: 43]. Such indices were categorised into: (i) metrics of structural complexity: indices reporting the length of units (sentences, T-units, clauses[2]), measured by counting the number of words; (ii) metrics of syntactic complexity: indices reflecting syntactic depth and dependency, that is, those based on coordination and subordination ratios as well as on clausal/T-unit embedding within other superordinate units; and (iii) metrics of categorial complexity: indices expressing the pervasiveness of nominal and verbal categories in the text.

At the second stage of the analysis, we carried out the qualitative analysis of the clausal and the phrasal complexity features, based on the taxonomy in Staples et al. [9]. The features are: sentence-final adverbial clauses of different types, *wh* complement clauses, verb + *that*-clauses, nouns, attributive adjectives, premodifying nouns and *of*-genitives. The analysis of such features required extensive manual disambiguation of the data examples.

[2] The notion of a T-unit is extensively used in complexity studies and is defined as "the shortest terminable units into which a connected discourse can be segmented without leaving any residue" [7: 34]. Bardovi-Harllg [8] notes that a T-unit normally comprises an independent along with its dependent clauses. For example, the expression *This would certainly continue to be the case with the CNT, but the UGT fared differently thanks to the support of the PSOE, its European partners and even the Spanish government, who had a strong interest in weakening the Communists* (CEH-2016-4) consists of one sentence, two T-units (*This would certainly continue to be the case with the CNT* and *the UGT fared differently thanks to the support of the PSOE, its European partners and even the Spanish government, who had a strong interest in weakening the Communists*) and three clauses (*This would certainly continue…, …but the UGT fared differently…* and *…who had a strong interest…*).

Table 2
L2SCA syntactic complexity indices

Structural complexity		MLS	mean length of sentence (no. of words)
		MLT	mean length of T-unit (no. of words)
		MLC	mean length of clause (no. of words)
Syntactic complexity	Coordination	CPC	coordinate-phrase/clause ratio
		CPT	coordinate-phrase/T-unit ratio
	Subordination	CS	clause/sentence ratio
		CT	clause/T-unit
		TS	T-unit/sentence ratio
		DCC	dependent-clause/clause ratio
		CTT	dependent-clause/T-unit ratio
Categorial complexity	Predicates	VPT	verb-phrase/T-unit ratio
	Nominals	CNT	complex-nominal/T-unit ratio
		CNC	complex-nominal/clause ratio

3. Results

The automated complexity indices are given in Table 3.

Table 3
L2SCA syntactic complexity indices in hard/soft sciences

Index	Hard sciences					Soft sciences				
	chemistry	physics	mathematics	engineering	*mean*	business	linguistics	history	political-sc	*mean*
MLS	32.3	26.26	27.99	27.34	*28.47*	32.68	31.47	63.9	35.84	*40.97*
MLT	29.75	25.35	25.87	25.33	*26.58*	30.87	29.04	56.74	31.88	*37.13*
MLC	20.03	16.33	15.12	15.49	*16.74*	17.65	16.52	29.42	16.02	*19.9*
CPC	0.49	0.31	0.17	0.34	*0.33*	0.66	0.41	0.37	0.28	*0.43*
CPT	0.74	0.47	0.29	0.52	*0.5*	0.88	0.7	0.71	0.56	*0.71*
CS	1.63	1.59	1.88	1.75	*1.71*	2.06	2.06	2.21	2.25	*2.14*
CT	1.5	1.19	1.74	1.62	*1.51*	1.79	1.84	1.93	2	*1.89*
TS	1.12	1.08	1.08	1.07	*1.09*	1.06	1.08	1.14	1.13	*1.1*
DCC	0.31	0.34	0.4	0.35	*0.35*	0.43	0.43	0.43	0.47	*0.44*
DCT	0.49	0.54	0.7	0.57	*0.58*	0.75	0.8	0.83	0.96	*0.84*
CTT	0.36	0.38	0.48	0.39	*0.4*	0.52	0.52	0.53	0.57	*0.53*
VPT	2.08	2.13	2.09	2.13	*2.11*	2.81	2.42	2.67	2.82	*2.68*
CNT	3.66	3.39	2.9	3.01	*3.06*	4.07	3.05	4.4	3.78	*3.83*
CNC	2.45	2.2	1.68	1.88	*2.05*	2.24	2.07	2.31	1.91	*2.13*

In an attempt to determine the relative weights of the complexity indices, a binomial linear regression analysis was applied to the data, implemented via the function ‘glm’ (‘stats’ package, R Core Team 2020). We operationalised a (backward-steps) reduction of the number of indices that led to the model in (1), with only the indices VPT (Verb phrases per T-unit), DCS (Dependent clause ratio), TS (T-unit/sentence ratio) and CPT (Coordinate phrases per T-unit). Both the C(oncordance) 0.918 and Nagelkerke R^2 0.653 discrimination indices indicate that the model is very good at explaining the variation.

(1) Definitive glm model ('***': 0,001, '*': 0,05)

```
            Estimate Std, Error z value Pr(>|z|)
(Intercept) -25,9115     4,0116  -6,459 1,05e-10 ***
vpt           3,4756     1,0531   3,300 0,000966 ***
dcs          10,6276     5,0567   2,102 0,035580 *
ts           10,1416     2,6373   3,845 0,000120 ***
cpt           3,8392     0,7312   5,250 1,52e-07 ***
```

Figure 1 presents the Random Forests (function 'cforest', 'party' package) corresponding to the model's fixed predictors, with an excellent C-index of 0.918. Figure 1 reflects the significant impact of the indices CPT, VPT and DCC on the variation hard/soft science, and the more minor contribution of TS to the model.

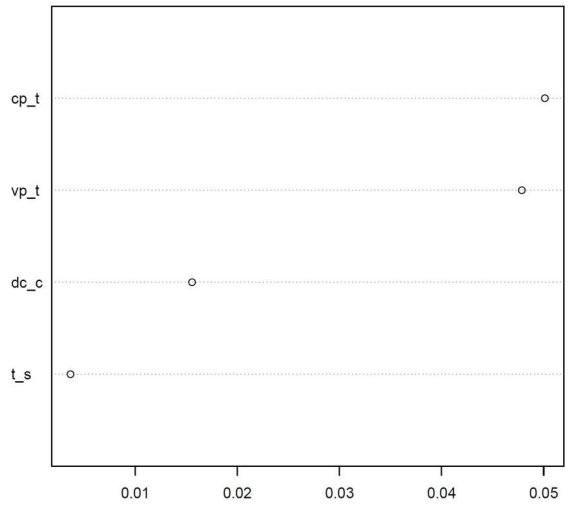


Figure 1: Dot chart of conditional variable importance

The interpretation of the findings revealed by the statistical analysis of the complexity indices per broad discipline, that is, hard and soft sciences, is as follows. The reduction of the indices led to a model with only 4 indices evincing different dimensions of linguistic complexity:

(i) syntactic complexity mirrored by pervasive coordination, as reflected by the index CPT, which calculates the ratio of coordinated phrases per T-unit

(ii) syntactic complexity determined by subordination within clausal units, as evinced by the index DCC, which expresses the amount of subordinate dependent clauses in matrix clauses, and in sentences, which has been corroborated by the statistical significance of the index TS, a telling indicator of the ratio of T-units per sentence

(iii) categorial complexity associated with the frequency of, specifically, verbal constituents in T-units, here captured by the index VPT.

Random Forests have demonstrated, on the one hand, that, out of the indices that proved to be very strong in the model, those measures evincing complexity triggered by coordination (CPT) and by the profusion of verbal categories (VPT), contribute to the variation of hard *versus* soft science to a greater extent than DCC and TS. On the other hand, the probability of higher values in the four complexity indices increases in academic writings categorised as soft science. In other words, greater ratios of coordination, subordination and the 'verby' status of texts can be taken as proxies for the categorisation of a research paper within the domain of social sciences and humanities. These results are in line with Biber et al, [10: 29] when they claim that "complexity is not a single unified construct, and it is therefore not reasonable to suppose that any single measure will adequately represent this construct". However, some remarks are in order here as regards the interpretation of our findings in light of the conclusions drawn by Biber and colleagues. In their multidimensional analysis of academic writing *versus* other more informal genres, Biber et al, [11] found that high(er) phrasal complexity and low(er) clausal complexity are characteristic features of academic English (as well as of newspaper and magazine writings). By contrast, the type of complexity evinced in personal,

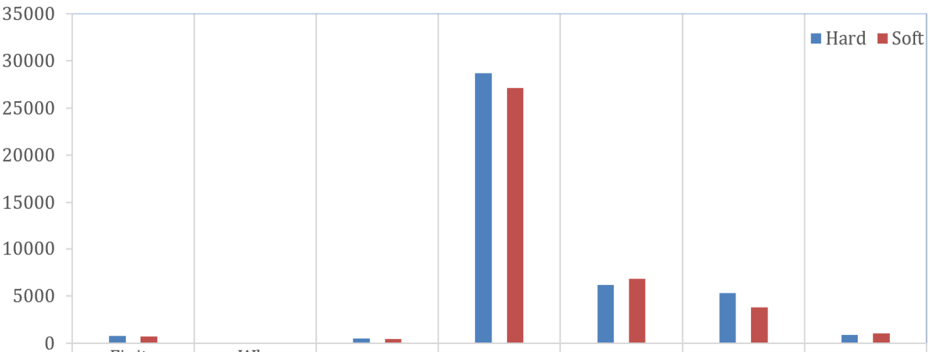
professional (even academic) spoken genres, as well as in popular written (novels, personal essays) discourse, is fundamentally clausal. Specifically, they contend that T-unit- and subordination-based (i,e, clausal) measures are not typical of academic writing but of conversational discourse, whereas nominal/prepositional (i,e, phrasal) measures are good indicators of academic writing. The statistical modeling of the complexity indices reported in this section has shown that subordination, coordination and the ‘verby’ status of sentences (or, better, T-units) are defining features of soft academic writing. As we see it, this conclusion does not invalidate a dominantly phrasal characterisation of academic writing when compared to more informal speech-based/related discourse, but gives support to the multifaceted nature of academic writing.

Subsequently, a more qualitative analysis of the frequencies of the features associated with clausal and phrasal complexity was carried out. The results of the such an analysis are shown in Figure 2, which provides the normalised frequencies (per 100,000 words) of the features.

All the differences in the use of the complexity features in hard and in soft sciences were found to be statistically significant at the level of 1%, except that of verb+*that*-clauses, which was significant at the 5% level. As can be seen in Figure 2, adverbial clauses were found to be more common in the corpus of the hard-science papers. A closer look at the types of adverbial clauses extensively employed in life and physical sciences revealed that the most frequently used one is the conditional clause, which accounts for almost a third of all adverbial clauses. This type of adverbial clauses is typically used in the comments for various calculations, formulas and theorems (see example 1). As regards the two features evincing complementation strategies, *wh*-clauses prevail in the soft research papers, whereas *that*-clauses are more frequent in the hard disciplines. Finally, the data demonstrates that, overall, phrasal complexity features, particularly, adjectival and prepositional phrases prevail in the soft-science texts, while nominal categories are more frequent in the hard sciences, particularly in chemistry, where they are used in long names of chemical entities and processes (see example 2).

(1) The next lemma expresses the important fact that if qC > 0 and if the excess measured relative to C is much smaller than the excess measured relative to pairs of planes with higher-dimensional axes… (JDG-2017-3).

(2) In addition, methyliminodiacetic acid (MIDA)-protected boronate esters were well tolerated (Chem-2016-4)



	Finite Adverbial clauses	Wh-complement clauses	Verb+ that-clauses	Nouns	Attributive adjectives	Premodifying nouns	Of-genitives
Hard	786,49	5,88	508,65	28674,29	6168,97	5304,27	862,92
Soft	720,50	25,29	471,04	27124,1	6808,53	3791,49	1062,76

Figure 2: Clausal/phrasal complexity features in hard/soft sciences

4. Conclusions

This study has tackled the analysis of linguistic complexity in professional academic writing in English. The analysis of automated indices of complexity in a corpus of research articles published in leading journals in hard (mathematics, chemistry, physics, engineering) and soft (linguistics, history, business, political science) science papers led to the following conclusions. Soft sciences demonstrate a significantly larger number of features associated with syntactic complexity, subordination and coordination ratios than the hard-science genre. The data have also revealed that the clausal-

complexity indices, in particular, the occurrence of sentence-final adverbial clauses, are significantly more frequent in the corpus of the hard-science papers. Phrasal complexity, measured here by the amount of adjectival and prepositional phrases, proved to prevail in the soft-science category, whereas the hard-science texts exhibited greater ratios of nominal categories.

An in-depth description of linguistic complexity in professional academic texts, along the lines of analyses of objectively depicted indices, can benefit the teaching of EAP/ESP writing in terms of guiding the production of discipline-specific language-learning materials that will address the needs of learners of different sciences in a more effective way. From the perspective of Data Driven Learning (DDL) approaches [12], EAP/ESP practitioners could employ teaching materials with examples from research papers in a particular discipline or group of disciplines (hard vs soft) with the purpose of helping students learn how to meet the necessary language and stylistic conventions established in a specific discipline. In this vein, concordance lines with the most common finite adverbial clauses could for example be employed to demonstrate the way in which clausal complexity is achieved and realised in hard sciences, while occurrences of adjectival and prepositional phrases from papers in soft disciplines would serve as an illustration of the type of phrasal complexity in this domain.

5. References

[1] C. Lambert, S. Nakamura, Proficiency-related variation in syntactic complexity: A study of English L1 and L2 oral descriptive discourse. International Journal of Applied Linguistics 29(2) (2019) 1–17. doi: 10.1111/ijal.12224

[2] J. E. Casal, J. J. Lee, Syntactic complexity and writing quality in assessed first year L2 writing. Journal of Second Language Writing 44 (2019) 51–62. doi: 10.1016/j.jslw.2019.03.005

[3] D. Mazgutova, J. Kormos, Syntactic and lexical development in an intensive English for Academic Purposes programme. Journal of Second Language Writing 29 (2015) 3–15. doi: 10.1016/j.jslw.2015.06.004

[4] K. Hyland, As can be seen: Lexical bundles and disciplinary variation. English for Specific Purposes 27(1) (2008) 4–21. doi: 10.1016/j.esp.2007.06.001

[5] R. F. Kelly-Laubscher, N. Muna, M. van der Merwe, Using the research article as a model for teaching laboratory report writing provides opportunities for development of genre awareness and adoption of new literacy practices. English for Specific Purposes 48 (2017) 1–16. doi: 10.1016/j.esp.2017.05.002

[6] X. Lu, A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. TESOL Quarterly 45(1) (2011) 36–62. doi: 10.5054/tq.2011.240859

[7] K. W. Hunt, Differences in grammatical structures written at three grade levels: The structures to be analysed by transformational methods. Report no. CRP-1998. Tallahasser: Florida State University, 1964.

[8] K. Bardovi-Harlig, A second look at T-unit analysis: Reconsidering the sentence. TESOL quarterly 26(2) (1992) 390–395. doi: 10.2307/3587016

[9] S. Staples, J. Egbert, D. Biber, B. Gray, Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. Written Communication 33(2) (2016) 149–183. doi: 10.1177%2F0741088316631527

[10] D. Biber, B. Gray, K. Poonpon, Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? TESOL Quarterly 45(1) (2011) 5–35. doi: 10.5054/tq.2011.244483

[11] D. Biber, B. Gray, K. Poonpon, Pay attention to the phrasal structures: Going beyond T-units – A response to WeiWei Yang. TESOL Quarterly 47(1) (2013) 192–201. doi: 10.1002/tesq.84

[12] T. F. Johns, Should you be persuaded: two samples of data-driven learning materials. English Language Research Journal 4 (1991) 1–16.