How complex is professional academic writing? A corpus-based analysis of research articles in 'hard' and 'soft' disciplines

Javier Pérez-Guerra[1]
Department of English
Universidade de Vigo
jperez@uvigo.es

Elizaveta A. Smirnova
Foreign Languages Department
HSE University / Universidade de Vigo
easmirnova@hse.ru

Abstract

This study focuses on the analysis of linguistic complexity in professional academic writing in light of the empirical evidence provided by a 1,597,000-word corpus of 'hard' (life and physical sciences) and 'soft' (arts and social) scientific research articles published in leading peer-review journals. Specifically, this investigation aims both to describe the complexity features of texts written by professional authors and to test the hypothesis that linguistic complexity varies across disciplines. Since previous studies have revealed that automatic complexity indices do not sufficiently succeed in providing a comprehensive description of complexity of texts, in this paper complexity has been measured in two ways: quantitatively through the indexes provided by Lu's (2010) L2 Syntactic Complexity Analyser, and through the more qualitative analysis of a selection of metrics associated with clausal and phrasal complexity in seminal studies. The data show, first, that syntactic complexity indices (basically, strategies of coordination and subordination) are statistically relevant to the characterisation of specifically the soft-science disciplines; second, that there is a continuum across subdisciplines within the broad distinction of soft versus hard genres; and, third, that the soft genre demonstrates a more stable productivity of clausal-complexity strategies, while phrasal-complexity features are more pervasive in the hard-science subcorpus.

Keywords: academic writing, complexity, corpus, hard sciences, soft sciences.

1 I am grateful to the Spanish State Research Agency and the European Regional Development Fund (MCIN/AEI/10.13039/501100011033 – PID2020-117541GB-I00), and Xunta de Galicia (grant no. ED431C 2021/52) for generous financial support.

Resumen

Este estudio se centra en el análisis de la complejidad lingüística en la escritura académica profesional a la luz de las pruebas empíricas aportadas por un corpus de 1.597.000 palabras de artículos de investigación científica '*hard*' (ciencias físicas y de la vida) y '*soft*' (humanidades y ciencias sociales) publicados en las principales revistas sometidas a revisión por pares. En concreto, esta investigación pretende describir las características de la complejidad de los textos escritos por autores/as profesionales y poner a prueba la hipótesis de que la complejidad lingüística varía según las disciplinas. Dado que estudios anteriores han revelado que los índices automáticos de complejidad no consiguen proporcionar una descripción exhaustiva de la complejidad de los textos, en este trabajo la complejidad se mide de dos maneras: cuantitativamente a través de los índices proporcionados por el L2 Syntactic Complexity Analyser de Lu (2010), así como a través de un análisis más cualitativo de una selección de métricas asociadas en estudios seminales a la complejidad clausal y sintagmática. Los datos muestran, en primer lugar, que los índices de complejidad sintáctica (básicamente, la coordinación y la subordinación) son estadísticamente relevantes para la caracterización de las disciplinas específicas de las ciencias '*soft*'. En segundo lugar, se demuestra que existe un continuo entre las subdisciplinas dentro de los géneros '*soft/hard*'. En tercer lugar, este estudio concluye que el discurso académico '*soft*' demuestra una productividad más estable de las estrategias de complejidad clausal, mientras que los rasgos de complejidad sintagmática son más dominantes en el subcorpus de las ciencias '*hard*'.

Palabras clave: escritura académica, complejidad, corpus, ciencias exactas, ciencias sociales.

1. Introduction

Different aspects of linguistic complexity have been explored from the perspective of corpus linguistics. As regards the complexity of writing, the literature has focused on the degree of complexity of L2 writing with respect to L1 writing (among others, Hinkel, 2003; Ai & Lu, 2013; Lambert & Nakamura, 2019), on correlations between text complexity, language proficiency and task types (Crossley & McNamara, 2012; Biber et al., 2016; Casal & Lee, 2019), or on the development of complexity in writing after intensive instruction (Crossley & McNamara, 2014; Mazgutova & Kormos, 2015). Despite the potential pedagogical implications of investigating the realisation of complexity features in expert writing, complexity in professional academic writing has been comparatively understudied to date. Research articles can be taken as a benchmark for optimal academic writing and can provide learners with "a rich and authentic introduction to the complexities and nuances of the genre" (Kelly-Laubscher

et al., 2017: 3). As claimed by Swales (1990: 177), research-oriented writing constitutes a core genre in academic discourse, comprising various subgenres such as dissertations, monographs and presentations. According to Casal et al. (2021: 2), the investigation of the degree of syntactic complexity in research articles can provide "important insights into the role that syntactically complex structures play in disciplinary RA [Research Article] writing practices". This is because it evinces the variation demonstrated by complex structures as demanded by their specific functional goals. It is from this angle that the study of complexity traits in research articles published in specialised peer-reviewed journals may help learners master the conventions of the genre.

This study undertakes a quantitative analysis of features associated with linguistic complexity, conducted on a 1,597,000-word corpus of research articles 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), published in leading journals. The goal is twofold: 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. Specifically, two research questions (RQ) are addressed here:

- RQ1. Which indices of linguistic complexity can serve as proxies for the characterisation of 'hard' and 'soft' sciences?

- RQ2. Do 'hard' and 'soft' scientific writings differ as regards the realisation of linguistic complexity features? Are these differences observable across specific disciplines?

This paper is organised as follows. Section 2 outlines the previous works on complexity in academic discourse and disciplinary variation. Section 3 describes the data and method of analysis. Section 4 presents the analysis and results of the study, which are summarised and discussed in Section 5.

2. Literature review

Linguistic complexity is a diverse notion whose scientific analysis has been carried out by exploring a number of lexical, structural and syntactic features (for example, Bulté & Housen, 2012) traditionally associated with production difficulty, proficiency and/or sophistication (Ortega, 2003: 492). This section summarises previous studies on linguistic complexity in academic writing specifically (Section 2.1) and informs of computational tools providing indices of complexity relying on taxonomies of linguistic features (Section 2.2).

2.1. Linguistic complexity in academic writing

Two dimensions can be identified in the study of linguistic complexity: phrasal and clausal complexity. Phrasal complexity is realised, for example, by the "dense use of embedded phrases functioning mostly as modifiers of a head noun", while clausal complexity is evinced by, for instance, the "dense use of dependent clauses functioning as clause constituents (complement clauses) or clausal modifiers (adverbial clauses)" (Biber & Gray, 2016: 141). In their seminal paper, Biber et al. (2011) measured complexity features in two corpora: a written corpus of research articles and a spoken corpus of face-to-face conversations. They concluded, on the one hand, that most of the measures evincing clausal subordination were more common in conversation than in academic writing. This conclusion is also in line with the findings presented in Biber et al. (1999). On the other hand, they demonstrated that academic writing featured more complex (specifically, noun) phrases. Therefore, it is suggested that grammatical complexity should be associated not with an extensive use of dependent clauses, typical of conversation, but with "linguistic units with phrases embedded in phrases" (Biber et al., 2021).

Gray (2015) explored linguistic complexity by paying attention to disciplinary variation. Her study focused on phrasal and clausal complexity in research articles in six disciplines, grouped into 'hard' sciences (physics and biology), social sciences (applied linguistics and political science) and humanities (history and philosophy). Gray concluded that clausal complexity is more prominent in the humanities and less salient in the hard sciences. This finding is in line with Staples et al.'s conclusion (2016), who studied (T-unit/clause-based) clausal and phrasal complexity in university students' writing in a number of disciplines. These authors found that "writers in arts and humanities disciplines and, to a lesser extent, the social sciences use more clausal features than writers in the life and physical sciences" (2016: 31). They also showed that, as university student writing develops, phrasal complexity increases and clausal complexity decreases. Gardner et al. (2019) clustered a number of linguistic complexity features in university student writing across four dimensions, and each dimension or cluster was explored according to the discipline, genre and level. Nesi & Gardner (2019) concluded that clausal complexity is more prominent in the so-called 'soft' disciplines and in the more conversational genres. In detail, in their study, complexity was shown to be achieved through the use of epistemic adverbials and stance nouns complemented by *that*-clauses in the texts in the soft sciences and by stance verbs in the conversational texts, such as narrative recounts. The authors contend that the findings might have important implications for the way in which academic writing is taught at universities.

Linguistic complexity in native and non-native academic writing has been the focus of a number of studies that justify this case study. To give a few examples, Gray

(2013) studied complexity variation in three types of research articles (theoretical, quantitative and qualitative) by investigating linguistic features related to elaboration/involvement *versus* informational density, one of the dimensions suggested by Biber already in his (1988) seminal study, and showed that the scope of complexity is broader than mere discipline variation since it also correlates with differences in the purpose and the type of evidence used in a particular study. Wu et al. (2020) explored syntactic complexity in research papers authored by ELF (English as a Lingua Franca) writers as opposed to those written by native speakers. By adopting Lu's (2010) indices, Wu et al. showed that ELF authors tend to use longer sentences, more coordinate phrases and complex nominals, and rely on nominal phrases to a greater extent than native writers. In a similar vein, Ruan (2018) explored phrasal complexity in journal abstracts by native English and non-native Chinese writers through the use of elaborated noun phrases (e.g. noun phrases with at least one pre-modifying element or a post-modifying prepositional phrase), and found that the non-native writers used more complex and elaborated noun phrases, particularly, employing significantly more noun premodifiers and multiple noun sequences, whereas the native writers opted for a more frequent use of the post-modifying *of* phrases. In their recent study, Yin et al. (2021) compared syntactic complexity indices provided by Lu's L2 Syntactic Complexity Analyser in emerging international research articles authored by L2 novice writers and those corresponding to publications by expert researchers. Their analysis revealed significant differences between the two types of texts; for instance, it was determined that novice writers use fewer verb phrases per T-unit and fewer instances of subordination, which might be explained by L1 transfer and by a lack of expertise in the use of such structures by the emerging writers.

2.2. Tools for measuring linguistic complexity

Over recent decades, several web-based tools have been developed for the automated analysis of the degree of complexity evinced by texts. To give a few examples, SyB http://sifnos.sfs.uni-tuebingen.de/SyB-0.1/#analyzer, developed at the university of Tübingen on the basis of the Common Text Analysis Platform (Chen & Meurers, 2016), provides 13 complexity indices of lexical, syntactic and discoursal phenomena. Coh-Metrix (http://tool.cohmetrix.com) detects basic cohesion, lexical, syntactic and semantic-discursive features, along with other metrics reporting textual lexical diversity and readability (approximately 200 metrics overall), which makes this tool especially useful for the study of text cohesion (as in, for example, Graesser et al., 2004; McNamara et al., 2010). The Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) (Kyle, 2016) measures a number of syntactic sophistication and complexity indices in learner writing. The indices include features already provided by Lu's (2010) Syntactic Complexity Analyser, discussed below, as well

as indices related to fine-grained clausal complexity, fine grained phrasal complexity and syntactic sophistication, employing the Stanford Neural Network Dependency Parser (version 3.5.1; Chen & Manning, 2014) and a Python XML parser that counts the relevant structures.

The tool used in this study is the L2 Syntactic Complexity Analyser (L2SCA henceforth; http://aihaiyang.com/software), developed by Lu (2010). It provides the frequencies and the ratios of complexity indices in Wolfe-Quintero et al. (1998) and Ortega (2003). L2SCA employs the Stanford parser (Klein & Manning, 2003) and is able to identify sentences, clauses, T-units[2], phrases, etc., using Tregex (Levy & Andrew, 2006), a utility that matches patterns in trees (https://nlp.stanford.edu/software/tregex.shtml). Lu tested the measures generated by the analyser in his (2017) study on syntactic complexity in L2 writing and found that a large number of the L2SCA metrics were "predictive of holistic measures of writing quality" (2017: 505). The indices of phrasal and clausal complexity that correlated with writing scores to a greater extent were: mean length of sentence, mean length of T-unit, mean length of clause, ratio of dependent clauses per clause and ratio of complex nominals per clause.

The measures provided by automated complexity analysers have been called into question in the literature. For example, Lambert & Kormos (2014: 2) contend that different types of subordinate constructions (e.g. adverbial clauses, complement

2 T-units, very similar to AS-units, are defined as the "shortest grammatically terminable units into which a connected discourse can be segmented without leaving any residue" (Hunt, 1964: 34), more specifically as "one main clause plus any subordinate clause or nonclausal structure that is attached to or embedded in it" (Hunt, 1970: 184; similarly in 1965: 36). 'Clause' is defined as "a structure with a subject and a finite verb (...), and includes independent clauses, adjective clauses, adverbial clauses, and nominal clauses" in Lu (2010: 481). To give an example, (i) illustrates a single T-unit (the whole utterance) consisting of one main clause (*I don't know ... to see something else*), which contains one embedded dependent clause (*why I was expecting...*), and the latter, in turn, includes the embedded dependent clause *to see something else*.

(i) I don't know [[why I was expecting [to see something else]].

The concept of c-unit, close to that of T-unit, also includes non-clausal structures which have communicative value, like *Coffee, please*. Lintunen & Mäkilä (2014: 394) develop their concept of U-unit, which is valid for the segmentation of spoken productions. In their words, a U-unit is "one independent clause or several coordinated independent clauses, with all dependent clauses or fragmental structures attached to it, separated from the surrounding speech by a pause of 1.5 seconds or more, or, especially in occurrences of coordination, a clear change in intonation and a pause of 0.5 seconds or more (depending on the average length of boundary pauses in the sample), containing one semantic unity" (2014: 385). Their analysis of the indexes of written and spoken L2 productions leads to the conclusion that "the choice of segmentation unit strongly affected the results, and that spoken language complexity may not be as different from written language complexity as it had been claimed in several earlier studies".

clauses controlled by verbs, complement clauses controlled by nouns) "emerge at different points in the developmental process [so,] the use of measures that conceptualise subordination as a unitary process can mask, rather than illuminate, developmental variation during task performance". They also point out that estimates of subordination can be inaccurate since, for example, clauses introduced by disjunctive markers such as *I see* or *I think* are parsed as subordinate by the analysers, whereas they do not necessarily illustrate syntactic subordination. In this vein, Wijers (2018) explored subordinate clauses in the writing of Dutch-speaking learners of Swedish as a foreign language and of native speakers, and concluded that the subordination ratios (the number of subordinate clauses divided by the total number of clauses, and the number of subordinate clauses divided by the number of T-units) are not efficient predictors of syntactic complexity since the differences in such ratios in learner and native writings were not significant, whilst the texts in which these subordination strategies were employed were clearly dissimilar. Similarly, Kyle & Crossley's (2018) in their study of TOEFL exams come to the conclusion that although the measures of phrasal complexity are better predictors of writing quality than clausal indexes, more fine-grained taxonomies of phrasal and clausal complexity per dependency/subordination type contribute to the explanatory power of the statistical model.

Some studies suggest that mode and task rank higher than proficiency as predictors of complexity. For instance, Biber et al.'s (2016: 662) study on complexity in spoken and written TOEFL exams demonstrated that "task-type differences on standardised language exams – associated with both speech versus writing and with different communicative purposes – are systematically associated with linguistic differences, especially with the use of grammatical complexity features". This finding is in line with previous research by the same team. In particular, Biber et al. (2011: 29; similarly, Biber & Gray, 2011) claimed that complexity cannot be seen as "a single unified construct", therefore, it is not reasonable to believe that any single measure will be able to "adequately represent this construct". They found 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.

As regards the influence of mode on the preference for the complexification strategies computed by the analysers, proficiency has been claimed not to run necessarily parallel to complexity, at least in the way the latter is measured by the analysers. In an influential study, Crossley et al. (2014) carried out a multifactorial analysis of L1 academic writing (argumentative essays), with such variables as score/grade, topic, writer's geographical area, timing and (handwritten or typed) production, as well as with the indexes provided by Coh-Metrix. It was found that the variable reflecting

the grade assigned to the essays explained only 5 percent of the total variance. Other studies give support to a certain parallelism between phrasal and clausal complexity. To give an example, with the objective of testing the measures provided by his analyser, Lu (2017: 505) operationalised syntactic complexity in L2 writing and concluded that "[a] large subset of the measures incorporated in L2SCA has been found to be predictive of holistic measures of writing quality as well. (...) [T]he following measures significantly correlated with writing scores: mean length of sentence, mean length of T-unit, mean length of clause, dependent clauses per clause, and complex nominals per clause", that is, indexes of phrasal and clausal complexity. By contrast, Lambert & Nakamura's (2019) study compared productions by L2 English produced by Japanese learners and native speakers through variables such as the proportion of simple, compound (coordination) and complex (nominal, adverbial and relative subordination) utterances, and phrasal-complexity measures like the ratio of words and modifiers per noun phrase. They prove the negative relationship between phrasal and clausal syntax: "[a]s the (...) measures of clausal complexity increase, the (...) measures of phrasal complexity tended to decrease and vice versa" (2019: 10), which is in keeping with the claims of Biber and colleagues. However, and this actually contradicts Biber's previous findings, they also observe a "negative relationship between proficiency and all [the] measures of phrasal syntax (words, modifier types, modifier tokens, and subordinated nouns per NP)": as proficiency increased, NPs became syntactically simpler, at least in the intermediate level of students' proficiency.

Substantial differences among disciplines and sub-registers or sub-genres have been reported in the literature as regards the use of complexity measures. Thus, Gardner et al. (2019: 670) analyse L1 university assignments and find that "the writing situation – disciplinary group [Arts and Humanities, Life Sciences, Physical Sciences and Social Sciences], genre family, discipline, and level of study – is key to interpreting each dimension" resulting from the application of the multidimensional analysis of 39 lexico-grammatical features associated with different aspects of linguistic complexity – see also Staples et al. (2016) in this respect.

Finally, in a recent paper on specifically developmental stages in L1 and L2 writing, Biber et al. (2020) argue that the omnibus complexity measures employed by automated complexity analysers fail to provide a comprehensive description of complexity, which must be compensated for by also describing the "multiple structural types, syntactic functions, and systematic patterns of variation across spoken and written registers" (2020: 13). In this vein, even though the current study does not investigate developmental stages associated with native academic writing, the analysis of the automated complexity indices will be enhanced by a more qualitative treatment of selected features evincing clausal and phrasal complexity.

3. Data and methodology

The analysis of linguistic complexity in professional academic writing has been conducted on a corpus of research papers published in peer-reviewed journals. Four disciplines fall into the category of so-called 'hard' science, and four into that of 'soft' science, thus representing a broad cross-section of academic discourse. The labels 'hard' and 'soft', commonly attributed to Storer (1967), are used to compare scientific fields on the basis of perceived methodological rigour, exactitude and objectivity. In a nutshell, the applied, empirical, experimental and natural disciplines (e.g. astronomy, biology, mathematics, physics) are considered 'hard', whereas the social sciences (e.g. history, linguistics, literature, sociology, political science) are categorised as 'soft'. Even though as stated by some researchers, the hard/soft division does not always adequately reflect the existing variation in the structure of knowledge in different disciplines (see, for example, Becher & Trowler, 2001; Nesi, 2002), this division can serve as a useful shorthand when attempting to describe the diversity of academic discourse (Dang, 2018). In this study, the hard-science subcorpus comprises articles in chemistry, physics, mathematics and engineering, and the soft-science subcorpus consists of texts in business studies, history, linguistics and political-science research articles. All the articles were published in leading academic journals, indexed in Scopus Quartile 1, in 2016-2020. This makes our corpus more up-to-date and homogeneous than, for instance, the Corpus of Academic Journal Articles (CAJA; Kosem, 2010), in which the category of journal articles also includes reports, reviews and progress reports published between 1993 and 2008. In our corpus we have aimed at warranting balance in terms of the number of tokens within each discipline. The texts were formatted for further textual analysis; for example, tables, formulas, graphs, charts, metadata and reference lists were removed from the documents. The size and details of the corpus are given in Table 1.

Table 1: Corpus

Discipline	No. texts	Word totals	*Journals*
HARD SCIENCES			
Chemistry	34	197,806	*Cell Chemical Biology (CCB)* *Chem* *Chemical Science (CS)* *Trends in Analytical Chemistry (TrAC)*
Physics	44	200,206	*Physics Letters B (PL)* *Reviews in Physics (RP)* *European Physical Journal C (EPJ)* *Nuclear Physics B (NPh)*
Mathematics	28	199,380	*Compositio Mathematica (CM)* *The Journal of Differential Geometry (JDG)* *Acta Mathematica (ActaM)* *Applied Mathematics and Computation (AMC)*
Engineering	34	198,926	*Automatica (Auto)* *Materials Characterisation (MC)* *International Journal of Engineering Science (IJES)* *Engineering (Engin)*
Totals	140	796,318	
SOFT SCIENCES			
Business	20	197,956	*The Journal of Management (JM)* *The Journal of Management Studies (JMS)* *Academy of Management Annals (AMA)* *Journal of Business Research (JBR)*
Linguistics	22	200,997	*Applied Linguistics (AL)* *Lingua (Ling)* *Modern Language Journal (MLJ)* *Language in Society (LS)*
History	21	199,394	*Contemporary European History (CEH)* *The Journal of Modern History (JMH)* *Journal of Global History (JGH)* *History of the Family (HF)*
Political science	25	202,040	*Political Analysis (PA)* *World Politics (WP)* *American Journal of Political Science (AJPS)* *British Journal of Political Science (BJPS)*
Totals	88	800,387	

This study undertakes 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 L2SCA (see Section 2.2). This software was chosen because, as stated in Lu (2010), the analyser's precision/recall rates and F-score are high, with an accuracy of 0.83+. Also, the complexity indices in L2SCA were identified specifically for the analysis of academic discourse, which makes them optimal for the purposes of this research. L2SCA provided the 14 indices given in Table 2 along with their descriptions, as in Lu (2011: 43). Such indices were categorised into:

i. metrics of structural complexity: indices reporting the length of units (sentences, T-units, clauses), 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

iii. metrics of categorial complexity: indices expressing the pervasiveness of nominal and verbal categories in the text, which have been identified as key measures of genre complexity in the literature (see Section 2.2).

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
		DCT	dependent-clause/T-unit ratio
		CTT	complex-T-unit/T-unit ratio
Categorial complexity	Predicates	VPT	verb-phrase/T-unit ratio
	Nominals	CNT	complex-nominal/T-unit ratio
		CNC	complex-nominal/clause ratio

The indices and the disciplines have been modelled statistically (see Section 4 for the detailed description of the analyses) by means of a number of multivariable methodologies (regression, Random Forests and clusters) in an attempt to both weigh the contribution of the individual complexity indices to the hard/soft distinction, and to determine similarities/differences among the hard and the soft disciplines as far as linguistic complexity is concerned.

Secondly, the corpus was tagged with TagAnt (Anthony, 2015), which employs the TreeTagger tagset (available at https://www.sketchengine.eu/english-treetagger-pipeline-2/) and processed with AntConc (Anthony, 2014), in order to carry out the qualitative analysis of the clausal and the phrasal complexity features in Table 3, based on the taxonomy in Staples et al. (2016).

Table 3: Clausal/phrasal complexity indices

Feature	Example
Clausal-complexity features	
Finite adverbial clauses of purpose introduced by the conjunctions *in order that*, *so that*	Moreover, p can be chosen *so that the next property is satisfied* (CM-2016-3).
Finite adverbial clauses of condition introduced by the conjunctions *if*, *unless*, *in the event that*, *provided that*	*Unless action was taken,* it might grow into a serious danger in a very short time (JMH-2016-2).
Finite adverbial clauses of concession introduced by the conjunctions *although*, *even though*, *despite the fact that*	*Although the above papers did not discuss switching speed,* typical integrated thermo-optical and electro-optical switching can reach GHz rates (RP-2016-5).
Finite adverbial clauses of time introduced by the conjunctions *after*, *before*, *when*, *until*, *as soon as*, *as*	No action could reasonably be taken *before the Western powers had given the signal* (CEH -2016-5).
Finite adverbial clauses of place introduced by the conjunction *where*	Once transcribed, the data are encoded with instances *where gestures…enacted by the right, left, and both hands being marked up* (AL-2016-2).
Finite adverbial clauses of reason introduced by the conjunctions *because*, *since*, *as*	*Since we observed a time-dependent accumulation of very long chain fatty acids,* we also investigated the expression levels of genes that showed significant accumulations during early necroptosis (CCB-2017-1).

Finite adverbial clauses of result introduced by the conjunction *so that*	The learning conditions were counterbalanced, *so that each participant learned half of the critical items in the WW condition and half in the ME condition* (AL-2016-4).
Finite adverbial clauses of manner introduced by the conjunctions *as if*, *as though*, *as*	In Western Europe the political discussions on Eureka were mostly conducted *as if the Soviet Union and the Cold War divide did not exist* (CEH-2016-2).
Finite adverbial clauses of contrast introduced by the conjunctions *while*, *whereas*	*While these adverbs have a basic restrictive function* which can be accounted for at the RL (Section 3.2.1.), they will be shown to have a number of different functions at the IL (Ling-2016-2).
Wh-complement clauses	The Chiefs of Staff of fronts and armies and scouting units do not know *where captives came from*...(JMH-2016-4).
Verb + *that*-clauses	Lagrangian is considered in the Einstein and in the Jordan frame, and we *demonstrated that several cosmological scenarios can be realised* (PL-2017-3).
Phrasal complexity features	
Nouns	Finally, phonetic *cues* as contrastive *stress* have been pointed out as another *factor* in determining the *interpretation* of *pronouns* (Ling-2016-1).
Attributive adjectives	The *crucial* role of *topological* defects was observed in a *new* type of phase transition in *two-dimensional* systems (PL-2016-1).
Premodifying nouns	Due to *space* limitations, we report the analyses including the six dummies for *company* level, but not including the statistical details of these dummies (JM-2016-3).
Of-genitives	In the absence *of enforceable global governance regimes*, the social responsibilities *of corporations* take on a new explicit political dimension (JMS-2016-2).

The qualitative analysis of the data required extensive manual pruning and the careful interpretation of the examples, for example, in order to determine the semantic

type of the adverbial clauses – to illustrate this, the conjunction *as* can introduce adverbial clauses of time (1), of reason (2) and of manner (3):

1. Moreover, *as the concentration of FabX was increased*, a second product appeared (CCB-2016-1)

2. *As the ammosamides share the same core structural features of lymphostin (Figure 1)*, they provide a unique opportunity to explore pyrroloquinoline alkaloid biosynthesis in a distinct genomic context (CCB-2016-2)

3. it should be possible to generate glycOMVs displaying a wide array of biomedically relevant glycotopes found on the surfaces of bacteria and human cells, *as we demonstrated here with T antigen and PSA* (CCB-2016-3)

4. Analysis of complexity

This section deals with the analysis of the complexity indices of the hard- and soft-science papers automatically generated by L2SCA (Section 4.1), and with the variation in the realisation of the syntactic complexity features in the same corpus (Section 4.2).

4.1. Automated complexity metrics

The mean values for each of the indexes per category (hard/soft) and discipline are given in Table 4.

Table 4: L2SCA syntactic complexity indices in hard/soft sciences

Index	Hard sciences					Soft sciences					Mean
	chemistry	physics	mathematics	engineering	*mean*	business	linguistics	history	political-sc	*mean*	
MLS	30.61	25.71	24.48	26.10	*26.72*	29.17	30.99	51.52	32.35	*36.06*	*31.43*
MLT	28.70	24.36	23.32	24.57	*25.23*	27.68	28.26	46.08	29.14	*32.83*	*29.06*
MLC	18.68	15.14	13.83	16.28	*15.98*	16.13	15.07	23.38	15.02	*17.42*	*16.70*
CPC	0.52	0.30	0.23	0.44	*0.37*	0.63	0.43	0.44	0.32	*0.45*	*0.41*
CPT	0.80	0.48	0.38	0.65	*0.58*	1.08	0.79	0.86	0.62	*0.84*	*0.71*
CS	1.65	1.72	1.76	1.61	*1.68*	1.83	2.07	2.22	2.16	*2.07*	*1.88*
CT	1.55	1.62	1.68	1.51	*1.59*	1.73	1.88	1.96	1.94	*1.88*	*1.74*
TS	1.07	1.06	1.04	1.06	*1.06*	1.06	1.10	1.13	1.11	*1.10*	*1.08*
DCC	0.34	0.36	0.40	0.32	*0.35*	0.41	0.44	0.44	0.46	*0.44*	*0.40*
DCT	0.54	0.60	0.68	0.50	*0.58*	0.73	0.85	0.89	0.91	*0.85*	*0.71*
CTT	0.39	0.41	0.46	0.36	*0.40*	0.50	0.54	0.54	0.56	*0.54*	*0.47*
VPT	2.17	2.17	2.22	2.01	*2.14*	2.50	2.69	2.72	2.71	*2.66*	*2.40*
CNT	3.82	3.37	2.92	3.24	*3.34*	3.9	3.89	4.76	3.85	*4.10*	*3.72*
CNC	2.47	2.10	1.74	2.18	*2.12*	2.27	2.08	2.44	2.00	*2.20*	*2.16*

In an attempt to determine the relative weights of the complexity indices within a multivariate model, a binomial linear regression analysis was applied to the data, implemented via the functions 'glm' in the 'stats' package (R Core Team, 2022).[3] More specifically, the model was intended to identify the complexity indices that have a significant determining effect on the categorisation of research writings into the hard- or the soft-science categories (i.e. the dependent variable). When the 14 complexity indices identified by L2SCA (Table 2) entered the model, the collinearity among indices revealed by the function 'vif' (Variance Inflation Factor, 'car' package; Fox & Weisberg, 2019) was severe. This statistical function indicates the presence in the model of predictors or variables (i.e. indices) that are not significantly informative because of mutual convergence with the values provided by other factors. This weakness of the model was overcome by operationalising the reduction of the number of indices in the initial model with no statistically significant loss of explanatory power. Firstly, taking backward steps (step(emode_glm, direction="backward"; 'MASS' package) caused such a harmless reduction, which led to a reduced model with only the L2SCA indices MLT, MLC, VPT, DCC, CS, CT and CPT. Secondly, the regression model (glm(formula = hardsoft ~ mlt + mlc + c_s + vp_t + c_t + dc_c + cp_t, family = binomial) revealed that the complexity indices MLC and MLT (Mean length of T-unit) did not contribute significantly to the characterisation of hard/soft academic writings, as shown in (4):

(4) Summary of glm model (significance conventions: '***': 0.001, '**': 0.01, '*': 0.05, '': 0.1)

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.7559     4.1081  -1.158 0.246996
mlc          -0.3761     0.2432  -1.546 0.122000
mlt           0.2336     0.1432   1.632 0.102762
dc_c         15.2052     4.0120   3.790 0.000151 ***
c_t         -12.2593     3.0563  -4.011 6.04e-05 ***
vp_t          3.1525     0.7440   4.237 2.27e-05 ***
c_s           5.2790     0.9986   5.287 1.25e-07 ***
cp_t          3.1392     0.5108   6.145 7.98e-10 ***
```

3 In an attempt to check the effect associated with the random variability of the indices per individual text, a generalised linear mixed-effects model was implemented on top of the fixed model, where the variable 'file', encoding for each of the 234 file names in the corpus, served as the random variable. For that purpose, we used the 'glmer' function in R (glmer(hardsoft ~ (1|file) + dc_c + c_t + vp_t + c_s + cp_t, data=data, family=binomial, control= glmerControl(optimiser="bobyqa")); 'lme4' package, Bates et al., 2015). The difference between the mixed- and the fixed-effects models' AICs (395.02 and 393.29, respectively) did not prove to be statisticaly significant (p=0.6072), which also leads to the conclusion that the mixed-effects model does not add explanatory force to the generalised linear model with the 4 fixed factors (indices).

In fact, dropping MLC and MLT from the dataset did not trigger a substantial statistical loss of the model's explanatory power, the difference between the models' AICs with and without MLC+MLT not being significant. As a consequence, we have opted for the simplest model summarised in (5), with only the indices VPT (Verb phrases per T-unit), DCC (Dependent clause ratio), CS (clause/sentence ratio), CPT (Coordinate phrases per T-unit) and CT (clause/T-unit).

(5) Summary of definitive glm model (significance conventions: '***': 0.001, '*': 0.05)

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.0875     1.2430  -8.920  < 2e-16 ***
dc_c         13.9837     3.8285   3.652  0.00026 ***
c_t          -8.1453     1.8276  -4.457 8.32e-06 ***
vp_t          3.3050     0.7395   4.469 7.84e-06 ***
c_s           5.1488     0.9829   5.239 1.62e-07 ***
cp_t          3.1483     0.4855   6.484 8.91e-11 ***
```

The VIF results are now low, ranging from 1.15 to 8.53, which reflects a lack of severe collinearity in the definitive model. Also, both the C(oncordance) 0.895 and Nagelkerke R^2 0.586 discrimination indices provided by 'lrm' ('rms' package; Harrell, 2021) indicate that the model is very good at explaining the variation (C>0.9 reveals the model's outstanding fit and predictive power, and R^2>0.4 it plausibility) and, consequently, adequate for the research question.

Logistic regression was used to assess the significance of the contribution of the indices to the overall categorisation of texts into hard and soft sciences. Random Forests (function 'cforest', 'party' package; Hothorn et al., 2006), first applied to linguistic analysis by Tagliamonte & Baayen (2012), rank predictors according to their impact on the explanation of the variation. Figure 1 presents the Random Forests corresponding to the model's fixed predictors. The goodness of fit of the model, as indicated by a C-index of 0.9487 computed by the function 'somers2' ('Hmisc' package; Harrell et al., 2020), is excellent and, as expected, slightly better than that of the regression model (C=0.895).

Figure 1: Dot chart of conditional variable importance

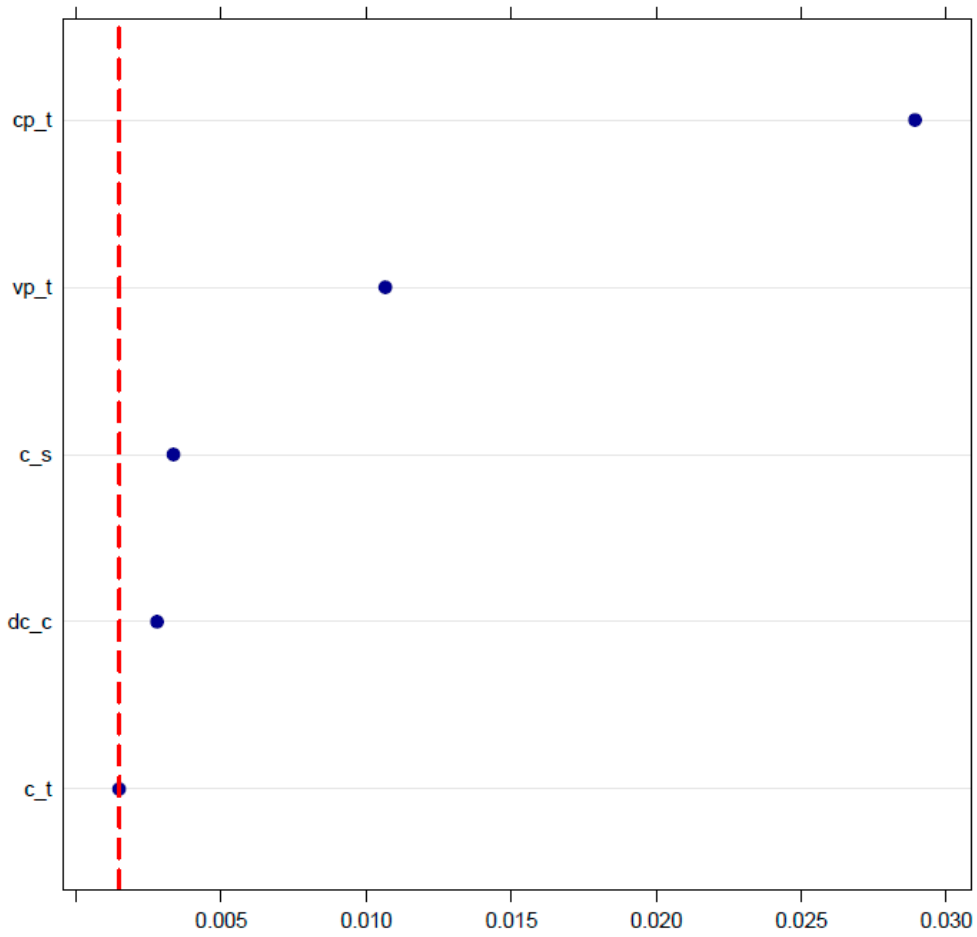
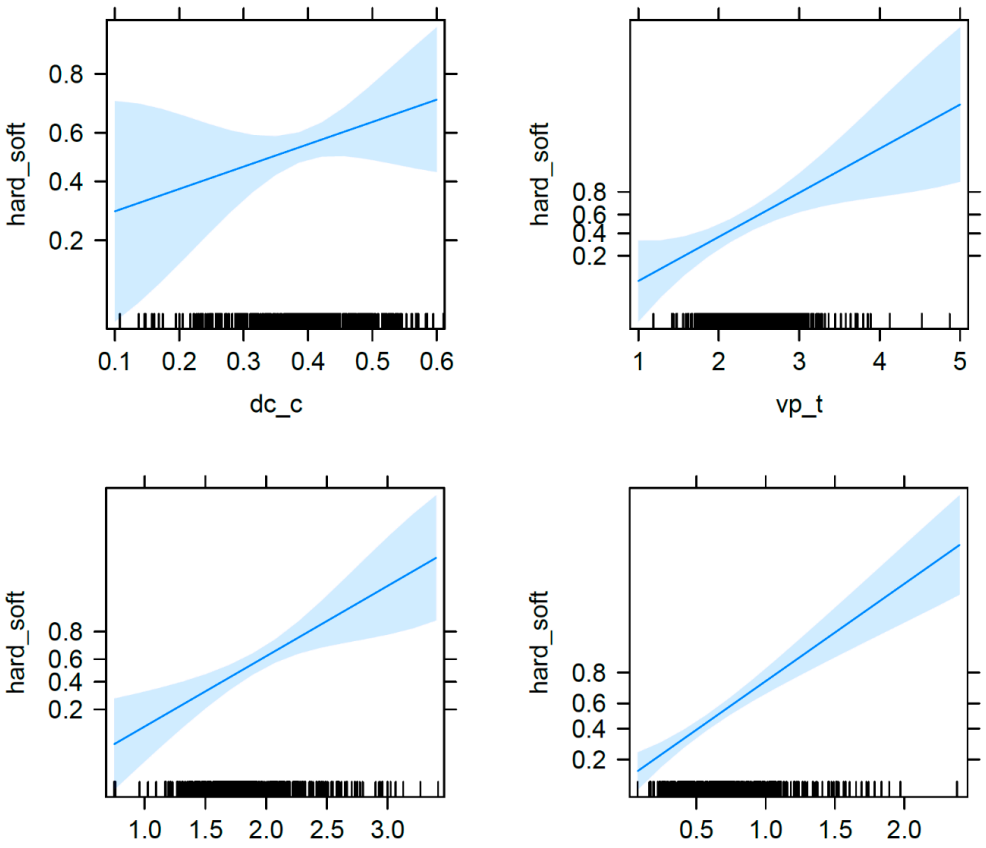


Figure 1 reflects, first, that the CT index exerts a very weak influence on the model, as indicated by its very low conditional importance value, and thus paves the way for the disposal of this variable. Second, Figure 1 evinces the significant impact of the indices CPT, VPT and DCC on the variation hard/soft science. The effects plots in Figure 2 provide a more detailed picture of the correlation between the significant indices and the categorisation of the research articles.

Figure 2: Effect plots



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. Firstly, as a consequence of the overlap of units such as sentence, clause and T-unit in specifically formal academic writing, the regression analysis showed severe collinearity among the initial 14 complexity indices. 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 number of subordinate dependent clauses in matrix clauses, and in sentences, which has been corroborated by the statistical significance of the index CS, a telling indicator of the ratio of clauses per sentence

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

Random Forests and the analysis of effects 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 CS. 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 give support to Biber et al. (2011: 29; similarly, Biber & Gray, 2011) 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. (1999, 2013) 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 modelling 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.

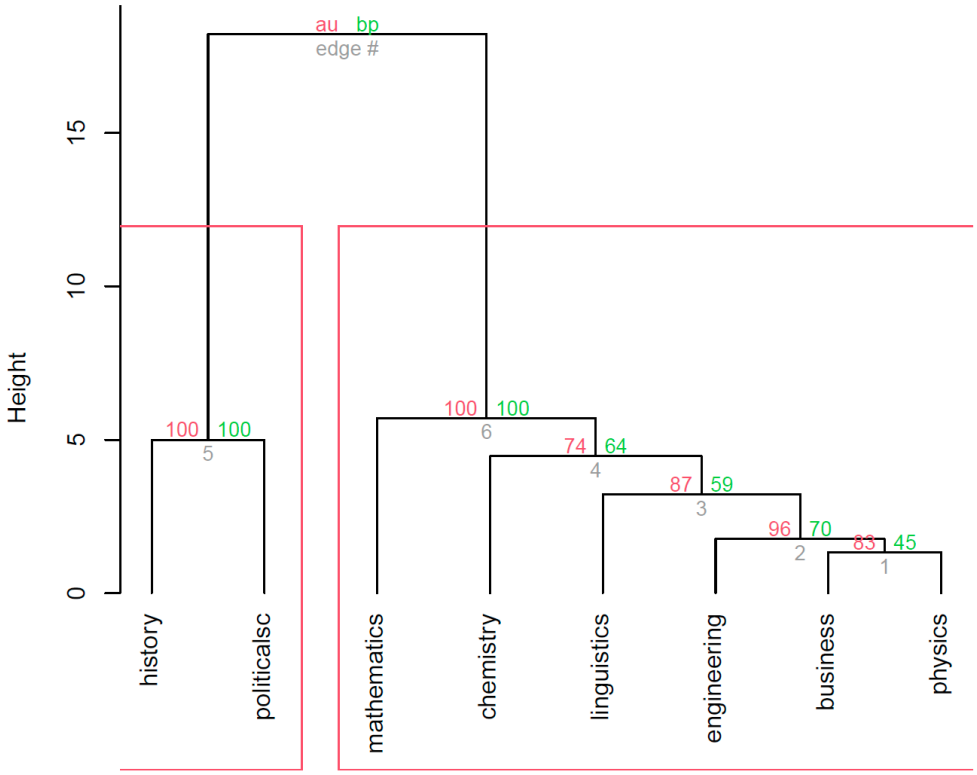
Our data confirm that, within the academic genre, the complexity strategies that serve to categorise soft- and hard-science articles are different, a continuum being observed between applied, empirical, experimental and natural disciplines, and social sciences and humanities as regards the productivity of, for example, coordination or subordination. This finding is in keeping with previous studies that highlighted substantial differences among academic disciplines, and sub-registers or sub-genres as regards the use of complexity measures. To give a few examples, Gardner et al. (2019: 670) analysed successful university student writing (about two thirds of which

was written by students who declared English as their first language) and found that "the writing situation – disciplinary group [arts and humanities, life sciences, physical sciences and social sciences], genre family, discipline, and level of study – is key to interpreting each dimension", resulting from the application of the multidimensional analysis of lexico-grammatical features associated with different aspects of linguistic complexity. Also, Hardy & Friginal (2016) confirmed that there is a continuum of academic paper types even within each of the dimensions recognised by Biber's multidimensional analysis. Our results also align with, for example, Nesi & Gardner's (2019), who, as already mentioned in Section 2.1, concluded that clausal complexity is more prominent in the so-called soft disciplines and in the more conversational genres[4].

A hierarchical agglomerative clustering algorithm was applied ('hclust' function, 'ward.D2' method, 'pvclust' package; Suzuki et al., 2019) to identify subgroups of disciplines based on the pervasiveness of all the complexity indices. The analysis used a Behavioural Profiles approach, in which the data are represented as vectors of proportions of each level of each variable ('bp' function, 'pvclust' package). With this technique, the numerical differences between vectors are operationalised as 'distances' ('dist' function, 'canberra' method, 'pvclust' package), which determine how the disciplines are grouped into clusters. The optimal number of clusters was calculated by means of the function 'silhouette' ('cluster' package; Maechler et al., 2019). The clusters are represented as tree leaves or branches of dendrograms (Levshina, 2015: 316), the most similar of which (i.e. those with the smallest 'distance') are merged together. Figure 3 displays a dendrogram of the clustering of the varieties.

4 The detected differences between subdisciplines are discussed in detail in Section 4.2.

Figure 3: Hierarchical cluster analysis of disciplines



The eight disciplines were grouped into the two statistically optimal clusters represented by the boxes in Figure 3. The stability of the clusters and their fit with the data was measured by the function 'pvclust' ('pvclust' package), which quantifies the uncertainty in the clusters by implementing multiscale bootstrap resampling to calculate the Approximately Unbiased (AU) *p*-values of each – the closer AU *p* is to 1, the greater the statistical significance of the cluster.

This technique identified two groups of disciplines: on the one hand, history and political science (both soft sciences), and all the remaining disciplines, on the other. To check the plausibility of such a grouping of disciplines, we carried out a qualitative analysis of textual samples randomly selected from the corpus of research papers. In detail, we extracted complete sentences, amounting to at least 150 words, from texts belonging to the disciplines of linguistics (in (6)) and business (in (7)), clustered together with the other hard-science texts, and of history (in (8)), a discipline included in the first (soft-science) cluster. The frequencies of finite verbal groups (in italics in

(6)-(8)), of finite subordinate clauses (starting with '[' in (6)-(8)) and of coordinating conjunctions (in boldface italics), have all been explored in the samples in an attempt to compute the most significant complexity indices of the hard *versus* soft distinction.

(6) [Because L2 proficiency *is* an important predictor of contextual vocabulary learning ***and*** only limited control over participants' proficiency *was* possible at recruitment, their L2 lexical proficiency *was* further estimated using the following published instruments: LexTALE *was* used as a measure of receptive vocabulary knowledge, and Laufer ***and*** Nation's (1999) vocabulary levels test of controlled-productive ability (PVLT) *was* used to measure their productive vocabulary knowledge (Table 1). PVLT *was* measured at the 2,000 ***and*** 5,000 word frequency levels, ***and*** the average score *was* used in the data analyses. Furthermore, [because larger working memory *tends* to positively correlate with word learning in L1 ***and*** L2, ***and*** [because both word-writing ***and*** meaning deliberation *may* consume the limited processing resources needed to create form-meaning associations (Barcroft 2006), participants' working memory *was* measured using an Operation Span (O-Span) task. Individual L2 vocabulary scores ***and*** working memory (O-Span) scores *were* included as covariates in the data analyses of the immediate ***and*** delayed tests. (AL2016-4)

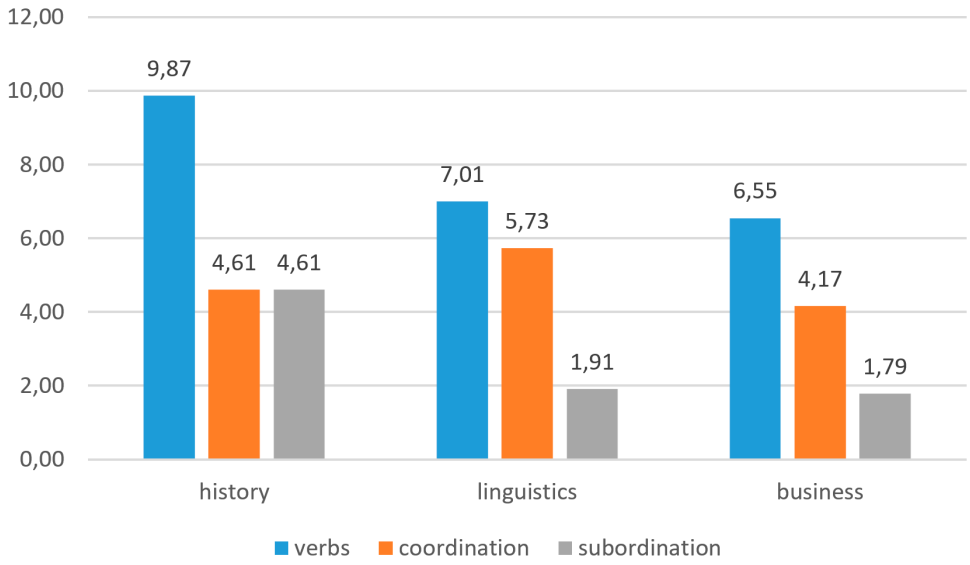
(7) In the previous section, we *have demonstrated* [that political CSR pursued by an organisation *is* related to individual-level CEO characteristics, i.e., their value orientation ***and*** subsequent behaviour. We *argued* [that CEOs with a social welfare orientation *are* likely to display an integrative responsible leadership style ***and*** *motivate* their organisations to engage in substantive political CSR (high-involvement MSI, second-order social innovation), [while CEOs with a strong sense of fiduciary duty *are* more likely to practice an instrumental responsible leadership style ***and*** *pursue* incremental political CSR (low-involvement MSI, first-order social innovation). We *will examine* in the following factors at the individual, organisational ***and*** societal level [that *moderate* the expected relationships between value orientations ***and*** leadership styles, emphasizing the relationship between social welfare orientation ***and*** an integrative responsible leadership style. Against the backdrop of political CSR we *discuss* exemplary, multilevel contingencies: at the individual level, the ability of CEOs to cope with complexity *is* a precondition to be able to respond to the complex institutional ***and*** relational environment of global business. (JMS2016-2)

(8) Apart from 'work on the party line' ***and*** similar tasks within the category of ideological work ***and*** propaganda, the committee *stressed* [that 'the entire work *is* focused on preparing ***and*** grooming the fighters for the struggle for national liberation'. A clandestine unit of 150 people *was created* ***and*** *organised* as a partisan military detachment, ready to join the partisans [when the moment *came*. It *was* clear [that these preparations *were* most relevant for young, able-bodied men ***and*** the few women

[who *could join* the partisan units in battle. [What *was* less clear *was* [what *was going to happen*, [when such opportunity *arose*, to the people unsuitable for partisan life: the sick, the elderly ***and*** children. The moment *came* soon enough. On 8 September 1943 the news of the armistice between Italy ***and*** the allies *reached* the camp. The exact chronology of events in the camp over the next few days *remains* murky. (CEH-2016-3)

The normalised frequencies of the three complexity strategies in the texts, displayed in Figure 4, corroborate the deviation of the linguistics and the business texts from the soft-science of history ones, the only exception to this clear-cut trend being the frequency of coordination in the linguistics extract, and, in consequence, the plausibility of the clusters sketched in Figure 3.

Figure 4: Complexity measure in subdiscipline samples



4.2. Syntactic complexity features

This subsection is devoted to a qualitative analysis of the frequencies of the features associated with clausal and phrasal complexity, as described in Section 3. Table 5 provides the raw and normalised frequencies (per 100,000 words) of the 16 clausal/phrasal complexity features as well as the measures of statistical significance (*p*-values resulting from chi-square test) of the variation hard- *versus* soft-sciences – statistical significance has been conventionalised as follows: '***' when $p \leq .001$ and '**' when $p \leq .01$.

Table 5: Clausal/phrasal complexity features in hard/soft sciences: raw, (normalised frequencies)

Feature	Hard sciences	Soft sciences	χ^2	*p*-value	
Finite adverbial clauses:					
Purpose	16 (2.01)	17 (2.13)	–	1	
Condition	1545 (194.10)	836 (104.50)	213.83	<0.00001	***
Concession	241 (30.28)	538 (67.25)	110.92	<0.00001	***
Time	733 (92.09)	1442 (180.25)	226.56	<0.00001	***
Place	1842 (231.41)	678 (84.75)	541.82	<0.00001	***
Reason	1195 (150.13)	1142 (142.75)	1.44	0.2309	
Result	176 (22.11)	46 (5.75)	75.61	<0.00001	***
Manner	113 (14.20)	198 (24.75)	22.26	<0.00001	***
Contrast	618 (77.64)	828 (103.50)	29.13	<0.00001	***
Adverbial clauses total	*6479 (813.94)*	*5725 (715.63)*	*50.00*	*<0.00001*	***
Wh-complement clauses	53 (6.66)	371 (46.38)	235.33	<0.00001	***
Verb+ *that*-clauses	3566 (447.99)	3964 (495.50)	18.87	0.000014	***
Nouns	224367 (28186.81)	219641 (27455.13)	60.43	<0.00001	***
Attributive adjectives	51917 (6522.24)	56727 (7090.88)	177.01	<0.00001	***
Premodifying nouns	40102 (5037.94)	30583 (3822.88)	1274.2	<0.00001	***
Of-genitives	6995 (878.77)	9219 (1152.38)	290.61	<0.00001	***

Most of the differences in the use of the complexity features in hard and in soft sciences are statistically significant. As for the clausal complexity features, overall, adverbial clauses were found to be more frequent in the corpus of the hard-science papers. Unlike in Staples et al. (2016), which examined finite adverbial clauses 'in bulk', Table 5 provides the frequencies of the different semantic types of adverbial clauses, which manifest significant differences across the two broad hard/soft categories. As regards the two features evincing complementation strategies, *wh*-clauses and *that*-clauses prevail in the soft research papers. Finally, the trends revealed by the data as far as phrasal complexity is concerned are, first, the preference for verbal, adjectival and prepositional phrases in the soft-science texts and, second, for nominal categories in the hard sciences.

Table 6 gives the distribution of the complexity features across individual disciplines, with raw and normalised frequencies per 100,000 words. To measure the dispersion of the complexity features in the corpus we used Juilland's D, which is considered to be the most reliable dispersion coefficient (Rayson, 2003: 94), ranging from 0 (a perfectly uneven distribution) to 1 (a perfectly even distribution). SD and Juilland's D were calculated for relevant frequency values.

Table 6: Clausal/phrasal complexity features across disciplines: raw, (normalised frequencies)

Feature	Hard sciences						Soft sciences					
	Chemistry	Engineering	Maths	Physics	SD	D-dispersion	Business	History	Linguistics	Political science	SD	D-dispersion
CLAUSAL COMPLEXITY												
Finite adverbial clauses:												
Purpose	2 (1.01)	4 (2.01)	3 (1.51)	7 (3.5)	0	0.59	6 (3.03)	6 (3.02)	4 (1.99)	1 (0.5)	0	0.59
Condition	64 (32.32)	300 (150.75)	954 (479.40)	227 (113.5)	0	0.44	143 (72.22)	176 (88.44)	175 (87.06)	342 (169.31)	0	0.60
Concession	109 (55.05)	65 (32.66)	25 (12.56)	42 (21)	0	0.59	209 (105.56)	104 (52.26)	96 (47.76)	129 (63.86)	0	0.66
Time	146 (73.74)	203 (102.01)	171 (85.93)	213 (106.5)	0	0.66	371 (187.37)	297 (149.25)	266 (132.34)	508 (251.49)	0	0.61
Place	120 (60.61)	384 (192.96)	694 (348.74)	654 (327)	0	0.63	117 (59.09)	130 (65.33)	189 (94.03)	242 (119.80)	0	0.63
Reason	176 (88.89)	198 (99.50)	565 (283.92)	256 (128)	0	0.37	271 (136.87)	183 (91.96)	334 (166.17)	354 (175.25)	0	0.63
Result	7 (3.54)	25 (12.56)	110 (55.28)	34 (17)	0	0.35	4 (2.02)	2 (1.01)	25 (12.44)	15 (7.43)	0	0.44
Manner	5 (2.53)	18 (9.05)	24 (12.06)	66 (33)	0	0.48	45 (22.73)	74 (37.19)	38 (18.91)	41 (20.3)	0	0.6
Contrast	159 (80.30)	165 (82.91)	64 (32.16)	230 (115)	0	0.55	187 (94.44)	161 (80.90)	232 (115.42)	248 (122.77)	0	0.65
Adverbial clauses total	*788 (397.98)*	*1362 (684.42)*	*2560 (1286.43)*	*1769 (884.5)*	*0*	*0.62*	*1353 (683.33)*	*1133 (569.35)*	*1359 (676.12)*	*1880 (930.69)*	*0*	*0.64*
Wh-complement clauses	14 (7.07)	10 (5.03)	13 (5.53)	16 (8)	0	0.64	185 (93.43)	28 (14.07)	78 (38.81)	80 (39.6)	0	0.58
Verb + *that*-clauses	628 (317.17)	811 (407.54)	1419 (713.07)	708 (354)	0	0.60	1054 (532.32)	848 (426.13)	952 (473.63)	1110 (549.5)	0	0.66
PHRASAL COMPLEXITY												
Nouns	58014 (29300.00)	57920 (29105.53)	51802 (26031.16)	56631 (28315.5)	0.12	0.67	58820 (29707.07)	49868 (25059.3)	52834 (26285.57)	58119 (28771.78)	0.12	0.66
Attributive adjectives	13902 (7021.21)	13189 (6627.64)	10579 (5316.08)	14247 (7123.5)	0.03	0.66	14696 (7422.22)	14512 (7.292.46)	12943 (6439.30)	14576 (7215.84)	0.03	0.66
Premodifying nouns	11023 (5567.18)	11460 (5758.79)	7448 (3742.71)	10171 (5085.5)	0.02	0.65	9375 (4734.85)	5646 (2837.19)	6902 (3433.83)	8660 (3936.36)	0.02	0.65
Of-genitives	1765 (894.41)	1725 (866.83)	1817 (913.07)	1688 (844)	0	0.67	2481 (1253.03)	2288 (1149.75)	2165 (1077.11)	2285 (1131.19)	0	0.67

Table 6 shows, on the one hand, that the distribution of phrasal complexity features is more even than that of clausal complexity features, which suggests that there is less disciplinary variation in their use. On the other hand, Table 6 reveals that the distribution of the clausal complexity features is more balanced in soft sciences than in the hard disciplines. The hard science that stands out in this respect is mathematics, with an extensive use of adverbial clauses of condition, place and reason. In the soft category, political-science texts noticeably demonstrate greater frequencies of adverbial clauses of time and reason.

Some final remarks are in order here concerning our fine-grained analysis of the automated indices per subdicipline and the, on occasions, intrinsic characteristics of academic writing. First, the frequencies for long sentences and for complex nominals, as identified by the analyser, were very salient in the chemistry texts, and this is partially due to the pervasiveness of outstandingly long names of chemical entities and processes in the field (see, in this respect, Dai et al., 2015), as illustrated in, respectively (9) and (10):

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

(10) We have demonstrated a new pathway of *unsaturated fatty acid synthesis* that is catalyzed by an enzyme, FabX, that has dual dehydrogenase/isomerase activities (CCB-2016-1)

Also, the coordination indices turned out to be outstanding in chemistry, when compared with the other hard sciences. This could be explained by the large number of descriptions of chemical experiments in the textual data, which usually involve several steps and deal with several chemical entities, as in (11).

(11) FabX was monitored at 280 nm *and* eluted at 15.26 min. (...) The standards were vitamin B12 (1.35 kDa), myoglobin (horse, 17 kDa), ovalbumin (chicken, 44 kDa), g-globulin (bovine, 158 kDa), *and* thyroglobulin (bovine, 670 kDa) (CCB-2016-1)

Another deviant hard-science subdiscipline was mathematics, where the high frequency of condition (12), place (13) and reason (14) adverbial clauses correlates with the latter's extensive use in the comments for calculations and formulas.

(12) *If we apply the recurrences σ1σ3 ... σn−1 to an arbitrary ca0, a1, ... ,an*, we obtain a linear combination of lower coefficients (CM-2016-2)

(13) For example, let G = SL4(C), and let w = w0s3, *where w0 is the longest element in W = S4, the symmetric group with four letters* (CM-2016-1)

(14) *Since the off-diagonal factors of R are all even in x0, x2, ... , xn and P (x) is even*, the diagonal factors of R(cid:91)R(cid:92) must be even as well (CM-2016-2)

Mathematicians' writing was characterised by Davis & Hersh (1981: 36) as giving "an impression that, from the stated definitions, the desired results follow infallibly by a purely mechanical procedure". Such an impression can be supported by the use of adverbial clauses, whose rhetorical function in mathematics is to explain calculations and formulas. Interestingly, time and reason adverbial clauses abounded also in the political-science research papers, specifically in comments for mathematical and statistical models (see examples (15) and (16)).

(15) Conversely, multiple imputation will offer small gains in bias reduction *when variables of theoretical interest have a low proportion of missing values* (PA-2016-1).

(16) Such bias inducers may not be troublesome in practice, however, either *because they can be identified for exclusion, as is sometimes the case for post treatment variables*, or *because the bias they induce tends to be small* (PA-2016-3).

Physics was found to employ a large number of attributive adjectives which, as in chemistry, are often part of terms used in the discipline. To give a few examples, *interatomic* bonds, *local* minimum, *magnetic* field, *massive* gravity, *black* hole.

5. Summary and conclusions

This study has looked at the linguistic complexity of professional academic writing by analysing automatically generated complexity measures and frequencies of complexity features in a corpus of research papers in four 'hard' and four 'soft' sciences. As regards the first research question 'do hard and soft disciplines differ as regards the selection of linguistic complexity indices or metrics?', the automated analysis of the data and their statistical modelling have shown that soft sciences demonstrate more signs of syntactic complexity, particularly of subordination and coordination ratios, than the hard-science genre. Also, the data have revealed the (statistically significant) pervasiveness of verbs in the soft academic writings, as compared to the hard-science texts. In response to the second research question 'do hard and soft scientific writings differ as regards the productivity of linguistic complexity features?', it has been found that the clausal-complexity indices, particularly, the amount of adverbial clauses, are more revealing in the corpus of the hard-science papers, where they also demonstrate a greater degree of variation within the category. As for phrasal complexity, this has been shown in the preference for verbal, adjectival and prepositional phrases in the soft-science texts, and for nominal categories in the hard sciences, where the latter often instantiate genre-specific terminology.

As far as the limitations of the current study are concerned, first, although the significance of our results has been statistically verified at all times, the limited size of the corpus suggests that the empirical results must be taken with a pinch of salt. Second, as already claimed by Hyland (2004: 30) and also corroborated by this investigation, the classification of sciences into hard *versus* soft is not able to capture disciplinary variation to the fullest. Therefore, to provide a fuller picture of the realisation of complexity features in academic discourse, additional sub-disciplines and sub-genres are needed.

All in all, despite the recognised limitations and differences among subdisciplines, we contend that the teaching of EAP/ESP writing will greatly benefit from the scientific study of linguistic complexity in academic genres, and that the rigorous description of the core complexity strategies adopted in professional academic writing will guide the production of discipline-specific language-learning materials that will effectively address the needs of learners of different sciences.

6. References

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In A. Díaz-Negrillo, N. Ballier & P. Thompson (eds), *Automatic treatment and analysis of learner corpus data* (pp. 249-264). Amsterdam: John Benjamins.

Anthony, L. (2014). AntConc (Version 3.4.4) [Computer Software]. Tokyo: Waseda University.

Anthony, L. (2015). TagAnt (Version 1.2.0) [Computer Software]. Tokyo: Waseda University.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Becher, T., & Trowler, P. R. (2001). *Academic tribes and territories* (2nd ed.). Philadelphia: Open University Press.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English*. London: Longman.

Biber, D. & Gray, B. (2011). Is conversation more grammatically complex than academic writing? In M. Konopka, J. Kubczak, Ch. Mair, F. Štícha & U.H. Waßner (eds), *Grammatik und Korpora 2009: Dritte Internationale Konferenz. Grammar & Corpora 2009: Third International Conference* (pp.7-61). Tübingen: Narr Verlag.

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge: Cambridge University Press.

Biber, D., Gray, B., & 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., Gray, B., & Poonpon, K. (2013). Pay attention to the phrasal structures: Going beyond T-units – A response to WeiWei Yang. *TESOL Quarterly*, *47*(1), 192-201.

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*(5), 639-668.

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, *46*, Article 100869.

Biber, D., Gray, B., Staples, S., & Egbert, J. (2021). *The register-functional approach to grammatical complexity: Theoretical foundation, descriptive research findings, application*. New York: Routledge.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & I. Vedder (eds), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21-46). Amsterdam: John Benjamins.

Casal, J.E., & Lee, J.J. (2019). Syntactic complexity and writing quality in assessed first year L2 writing. *Journal of Second Language Writing*, *44*, 51-62.

Casal, J.E., Lu, X., Qiu, X., Wang, Y., & Zhang, G. (2021). Syntactic complexity across academic research article part-genres: A cross-disciplinary perspective. *Journal of English for Academic Purposes*, *52*, Article 100996.

Chen, D., & Manning, C.D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 740-750). Doha, Qatar.

Chen, X., & Meurers, D. (2016). CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity at COLING, Osaka, Japan, 11th December.* (pp. 113-119). Osaka, Japan: The International Committee on Computational Linguistics.

Crossley, S.A., & McNamara, D.S. (2012). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading*, *35*(2), 115-135.

Crossley, S.A., & McNamara, D.S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*, 66-79.

Crossley, S.A., Allen, L.K., & McNamara, D.S. (2014). A Multi-Dimensional Analysis of essay writing. What linguistic features tell us about situational parameters and the effects of language functions on judgements of quality. In T. Berber Sardinha & M. Veirano Pinto (eds), *Multi-Dimensional Analysis, 25 years on: A tribute to Douglas Biber* (pp.197-238). Amsterdam: John Benjamins.

Dai, H.J., Lai, P.T., Chang, Y.C., & Tsai, R. T.H. (2015). Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenisation. *Journal of Cheminformatics*, *7*, 1-14.

Dang, T.N.Y. (2018). The nature of vocabulary in academic speech of hard and soft-sciences. *English for Specific Purposes*, *51*, 69-83.

Davis, P.J., & Hersh, R. (1981). *The mathematical experience*. Boston: Birkhauser.

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression*. Thousand Oaks: Sage.

Gardner, S., Nesi, H., & Biber, D. (2019). Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. *Applied Linguistics*, *40*(4), 646-674.

Gray, B. (2013). More than discipline: Uncovering multi-dimensional patterns of variation in academic research articles. *Corpora*, *8*(2), 153-181.

Gray, B. (2015). On the complexity of academic writing: Disciplinary variation and structural complexity. In V. Cortes & E. Csomay (eds), *Corpus-based research in applied linguistics: Studies in honor of Doug Biber* (pp. 49-78). Amsterdam: John Benjamins.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193-202.

Hardy, J.A., & Friginal, E. (2016). Genre variation in student writing: A Multi-Dimensional Analysis. *Journal of English for Academic Purposes*, *22*, 119-131.

Harrell, F.E.Jr. (2021). Regression Modeling Strategies. https://github.com/harrelfe/rms

Harrell, F.E.Jr. with contributions from Charles, D. et al. (2020). Hmisc version 4.3-1. https://CRAN.R-project.org/package=Hmisc

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly, 37*(2), 275-301.

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. (2006). Survival ensembles. *Biostatistics*, *7*(3), 355-373.

Hunt, K.W. (1964). *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.

Hunt, K.W. (1965). *Grammatical structures written at three grade levels*. NCTE Research Report No. 3. Champaign, IL: National Council of Teachers of English.

Hunt, K.W. (1970). Recent measures in syntactic development. In M. Lester (ed), *Readings in applied transformational grammar* (pp. 179-192). New York: Holt, Rinehert and Winston.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor, MI: University of Michigan Press.

Kelly-Laubscher, R. F., Muna, N., & van der Merwe, M. (2017). 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*, 1-16.

Klein, D., & Manning, Ch.D. (2003). Fast exact inference with a factored model for Natural Language Parsing. In S. Becker, S. Thrun & K. Obermayer (eds), *Advances in neural information processing systems* (pp. 3-10). Cambridge, MA: MIT Press.

Kosem, I. (2010). *Designing a model for a corpus-driven dictionary of academic English*. PhD, Aston University.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. PhD, Georgia State University, Atlanta, GA.

Kyle, K., & Crossley, S.A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal, 102*(2), 333–349.

Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics*, *35*(5), 607-614.

Lambert, C., & Nakamura, S. (2019). Proficiency-related variation in syntactic complexity: A study of English L1 and L2 oral descriptive discourse. *International Journal of Applied Linguistics*, *29*(2), 1-17.

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins.

Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. (pp. 2231-2234). Genoa: ELRA.

Lintunen, P., & Mäkilä, M. (2014). Measuring syntactic complexity in spoken and written learner language: Comparing the incomparable? *Research in Language, 12*(4), 377-399.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36-62.

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, *34*(4), 493-511.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik Maechler, K. (2019). cluster: Cluster analysis basics and extensions. R package version 2.1.0.

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*, *29*, 3-15.

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., & Graesser, A.C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*(4), 292-330.

Nesi, H. (2002). An English spoken academic wordlist. In A. Braasch & C. Povlsen (eds) *Proceedings of the Tenth EURALEX International Congress. Vol. 1*. (pp. 351-358). Copenhagen.

Nesi, H., & Gardner, S. (2019). Complex, but in what way? A step towards greater understanding of academic writing proficiency. In C. Danjo, I. Meddegama, D. O'Brien, J. Prudhoe, L. Walz & R. Wicaksono (eds.), *Online Proceedings of the 51st Annual Meeting of the British Association for Applied Linguistics: Taking Risks in Applied Linguistics, 6-8 September, 2018*. https://custom.cvent.com/01664CE00C344F7BA62E39C4CFE91FA8/files/0f77de05eb81461a8037170680562243.pdf (accessed on 22.06.2019)

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*, 492-518.

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. PhD thesis, Lancaster University. https://eprints.lancs.ac.uk/id/eprint/12287/1/phd2003.pdf (accessed on 20.12.2020)

Ruan, Z. (2018). Structural compression in academic writing: An English-Chinese comparison study of complex noun phrases in research article abstracts. *Journal of English for Academic Purposes*, *36*, 37-47.

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33*(2), 149-183.

Storer, N.W. (1967). The hard sciences and the soft: Some sociological observations. *Bulletin of the Medical Library Association*, *55*(1), 75-84.

Suzuki, R., Terada, Y., & Shimodaira, H. (2019). pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. R package version 2.2-0.

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

Tagliamonte, S.A., & Baayen, R.H. (2012). Models, forests and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change*, *24*, 135-178.

Wijers, M. (2018). The role of variation in L2 syntactic complexity: A case study on subordinate clauses in Swedish as a foreign language. *Nordic Journal of Linguistics*, *41*(1), 75-116.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawaii Press.

Wu, X., Mauranen, A., & Lei, L. (2020). Syntactic complexity in English as a lingua franca academic writing. *Journal of English for Academic Purposes*, *43*, Article 100798.

Yin, S., Gao, Y., & Lu, X. (2021). Syntactic complexity of research article part-genres: Differences between emerging and expert international publication writers. *System*, *97*, Article 102427.