International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 7, Issue.6, Nov-Dec 2024, page no. 299-308

To cite this article: Christoph Haase (2024). TEACHING LEXICO-SEMANTIC COMPLEXITY IN STUDENT ACADEMIC WRITING: CORPORA AND CORPUS TOOLS, International Journal of Education and Social Science Research (IJESSR) 7 (6): 299-308 Article No. 1012, Sub Id 1591

TEACHING LEXICO-SEMANTIC COMPLEXITY IN STUDENT ACADEMIC WRITING: CORPORA AND CORPUS TOOLS

Christoph Haase

Purkinye University, Usti nad Labem, Czech Republic

DOI: https://doi.org/10.37500/IJESSR.2024.7622

ABSTRACT

This contribution surveys methods to enhance the teaching of English for Academic Purposes via computer-aided corpus methods. It introduces data generated from a custom-made algorithm to automatically assign a score for the lexico-semantic complexity of a given text. It also presents findings from corpus studies that investigate the linguistic parameters of academic texts. The tool utilizes the WordNet project, which is part of the semantic web initiative. The texts analyzed in the study come from the self-compiled corpus project CUJOE which includes various registers of student academic writing, in contrast to other corpus data from specified academic texts and popular science articles. The article involves suggestions for using the tool to promote self-assessment of semantic complexity in the teaching of academic writing.

KEYWORDS: Academic English, English for Academic Purposes, Teaching Academic Writing, Learner Corpora, Corpus Analysis

1. INTRODUCTION

Corpora have been in use for teaching for a substantial number of years with specified applications generating research interest beginning in the early days of ReCall (see §1.2) and culminating in big-data supported AI tools like Grok, Gemini or ChatGPT. While these applications cover the side of content generation, it remains an opaque instrument and, for the learner, may provide a short-term solution to a problem (a due paper, a graded essay) but do little for enabling the learner to create and reflect. However, self-assessment is key for skill development. Multiple research outputs demonstrate the potential benefits of corpus-based methods in various areas of language teaching, learning, and research. For example on the effectiveness of corpus tools in learning and teaching, Cheng, Greaves, and Warren write about their findings that they "support the effectiveness of corpus tools in second language vocabulary learning, and provide important implications for the design and use of such tools in language teaching and learning contexts" (2021, p. 10). On the assessment side, Fawcett and Kuo maintain the "potential of using corpus-based approaches to inform language assessment practices, allowing for more valid and reliable measures of language proficiency" (Fawcett & Kuo, 2021, p. 177).

The following contribution presents the development and implementation of a corpus-based tool for the study and teaching of English for Academic Purposes (henceforth: EAP). The goals of this tool are twofold: firstly, to demonstrate that modern corpora can enhance the methods employed by educators at all levels, and secondly, to explain that these corpora are not static collections for simple look-up tasks but can serve as a backbone for more dynamic and engaging teaching methods when used in conjunction with computational tools. As Boulton and Cobb (2021) maintain, "academic vocabulary resource[s] can have a positive impact on L2 learners 'vocabulary knowledge and use, and that its effectiveness can be further enhanced by integrating it into classroom teaching and learning." (p. 8).

1.1 What corpora contribute

Corpora provide a valuable perspective for educators and learners that extends beyond the limitations of impressionistic analysis. They enable second-language speakers of English to learn from authentic examples and collocations, provided that the corpus is representative and avoids perpetuating language impenetrability. Learner corpora are generally used for data-driven or inductive learning rather than deductive, rule-based learning. Several publications focused on genre analysis using corpus-based methods. Bondi and Scott (2021) conclude that "corpus-informed genre analysis can help researchers gain a better understanding of the communicative features and rhetorical structures of academic genres" (p. 12). At the same time, an appropriate mix of approaches taken from both corpus and discourse analysis allow a better grasp of the materials for practitioners and learners alike by adding "the value of combining corpus and discourse approaches in the analysis of academic genres, and provides a nuanced understanding of the rhetorical features and functions (ibid).

The following paragraphs explore the relationship between the use of corpora and teaching, surveying research conducted in this field. The corpus used for the present study is described in detail, and the core of the paper presents a novel tool named ComplexAna, which has a significant impact on corpus research and education.

1.2 Corpora and language pedagogy

Numerous publications in the last two decades have shown that the use of corpora for teaching methodology is a relatively recent phenomenon. Early notable examples include Aston, Bernardini & Stewart (2001), Braun, Kohn & Mukherjee (eds.) (2005), Granger in Connor & Upton (eds.) (2004), Nesselhauf (2005), Renouf (ed.) (2006), Sinclair (2004), and as an initial spark, the ReCall 19, 3 (2007) special issue on "Incorporating Corpora in Language Learning and Teaching" edited by Chambers & Thompson. In the subsequent years, pedagogical interest in applying corpus results to teaching, particularly in ESL writing, has grown significantly. Corpora provide authentic materials and corpus-based studies are particularly useful for teaching reading and writing skills and the development of academic literacy. As Cheng, Greaves, & Warren note, "corpus tools have a positive and significant impact on learners' vocabulary knowledge and retention" (2021, p. 8). Fawcett & Kuo add that corpora

can change the face of language assessment: "corpora enable the assessment of language use that is more authentic, valid, and reliable than traditional assessments" (2021, p. 165).

However, few special uses of corpora exist for teaching highly specific and advanced forms of language communication. This project seeks to open new pathways away from corpora as a mere example-generating machine towards a more dynamic approach, introducing the SPACE corpus (published in Haase 2009, 2013a, and 2014), which stands for "Corpus of Specialized and Popular Academic English".

2. The CUJOE corpus

Development of CUJOE started in two steps. The first step was the establishment of SPACE (see below), the second step was the development of a corpus tool for lexico-semantic complexity called ComplexAna. In the following, the initial corpus for the study and teaching of academic vocabulary will be sketched.

2.1 Project rationale

The forerunner of the Corpus of UJEP students of English (CUJOE) is the SPACE corpus which aimed to provide a mid-scale yet comprehensive representation of the major branches of science, ranging from abstract quantum theory to concrete fields such as zoology and plant science. Initially, copyright issues posed a challenge, requiring contact with multiple individual authors or copyright holders. However, the emergence of pre-print servers provided a solution. The primary goal of this corpus was to generate interest in academic writing and address the observed need at the university level for high-quality writing courses that could encompass the highest levels of academic writing. This skill could not be taught by general educators because the scientific disciplines were too diverse, making a writing centre a necessity. As a result, a corpus was designed to offer a span of very different branches within the natural sciences. Further details on the SPACE corpus can be found in Haase 2009.

2.2 Compilation: Academic disciplines

The selection of examples plays a crucial role in addressing the perceived need of learners to practice academic writing. Different disciplines have varying demands for writing, with those in humanities and social sciences differing from those in engineering. In order to provide comprehensive coverage, it was decided to include a broad spectrum of examples. This is reflected in the double backbone of the SPACE corpus, which consists of texts from both physical sciences and biosciences. These texts were sourced from pre-print servers like arxiv, as well as openly accessible research results published in the Proceedings of the National Academy of Sciences (PNAS). Table 1 provides a summary of the corpus composition.

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 8, Issue.1, Jan-Feb 2025, p no. 299-308

Table 1: SPACE domains

subcorpus	descriptors	word count
arXiv	physics, astrophysics, quantum mechanics	809,320
New Scientist – physics	physics, astrophysics, computer science, quantum mechanics	203,470
Proceedings of the National Academy of Science (PNAS)	biochemistry, genetics, genetic engineering, microbiology	267,105
New Scientist - biosciences	biochemistry, genetics, genetic engineering, microbiology	30,499
Public Library of Science – Medicine (PLoS),	medicine, virology, clinical psychology, public health	217,254
New Scientist – medicine	medicine, virology, clinical psychology, public health	17,050
total		**1,544,149**

Many research findings obtained in years of empirical research on the SPACE corpus led to new approaches to the teaching of academic writing at university level. The subsequent step therefore was to give students the tools they needed to enhance their own academic output. The logical consequence was twofold, establishment of the CUJOE corpus and creation of a self-assessment tool.

2.3 CUJOE Sampling

CUJOE, the Corpus of UJEP Students of English, described first in Haase, 2019, represents an outgrowth of the findings from SPACE and compiles texts selected to meet the following criteria. The main parameters are summarized in Table 2.

Table 2: CUJOE parameters

Shared features		Variable features	
Age	*20-30*	Sex	*75%F, 25%M*
Learning context	*Degree in English*	Mother tongue	*Czech*
Level	*BA, MA*	Region	*Northern Bohemia*
Medium	*Written*	Other foreign languages	*diverse*
Genre	*Linguistics*	Practical experience	*diverse*
Technicality	*Digital*	Topic	*English language linguistics*
		Task setting	*assigned qualification*

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 8, Issue.1, Jan-Feb 2025, p no. 299-308

The parameters follow the design of ICLE, the International Corpus of Learner English (see Granger 1998, 2009). From the beginning, the corpus objective saw students as practitioners of the sociolect of academic English, thus as learners of this variety they become producers of learner English. With academic English as a variety, it represents therefore a helpful resource especially given that students can test out their own lexico-semantic proficiency in academic vocabulary under usage for the corpus tool ComplexAna. As no one is a ‚native speaker ‘of academic English, the difference between native and non-native practitioners is revealing. Chen & Baker write about their own research "The results of the study showed that Chinese students used fewer phraseological patterns in their writing than British students, and the phraseological patterns they did use were less complex. This indicates a need for EAP instruction to focus on phraseological patterns to improve the quality of Chinese students' academic writing." (2021, p. 11). This is supported by our own experiences with academic practitioners of the first language of Czech. The data obtained from CUJOE therefore reveal interesting patterns that can be related to the non-native nature of the academic output in student academic writing.

3. ComplexAna

3.1 Development

If an individual, whether a learner or a human analyst, is tasked with explaining the process of transfer mentioned previously, they would likely describe it as the movement of highly specialized words for specific objects and events down a scale of lexical specificity until a shared semantic core is achieved. This core may not be situated at basic-level categories (as would be suitable for a young audience), nor at a level requiring specialized knowledge in the field. This concept can be demonstrated using terms extracted from two comparable texts (corpus codes 0066PN and 0066NS), as shown in Table 3.

Table 3: Lexico-semantic complexity in two parallel examples

Academic text 0066PN	Popular-academic text 0066NS
biotic evolution	wiped out three-quarters of life
stomatal index	sparked a dramatic change
postboundary pCO2 rise	sudden greenhouse effect
multimillennial pCO2 perturbation	huge asteroid
stratigraphically well-dated	within 10,000 years of the impact

Assuming that the complexity of academic vocabulary is a marker of argumentative strength in an academic text, we can use it to systematize a lexico-semantic function and automatically profile texts for learners. This can be helpful in comparing texts and measuring their difficulty, as well as obtaining learner data from recognition tests to correlate with words that are considered highly specialized. Learners can then consciously employ the process of transfer to rephrase a given text, guided in both upward and downward directions of specialization. To make this feasible, we used WordNet, a

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 8, Issue.1, Jan-Feb 2025, p no. 299-308

linguistic ontology hosted at Princeton University (cf. Fellbaum, 1998), which disambiguates words into their superordinate and subordinate categories to create a network. The position of a lexical item in that network scales its ontological depth, which can be used to run statistics on the scaling of items in given texts.

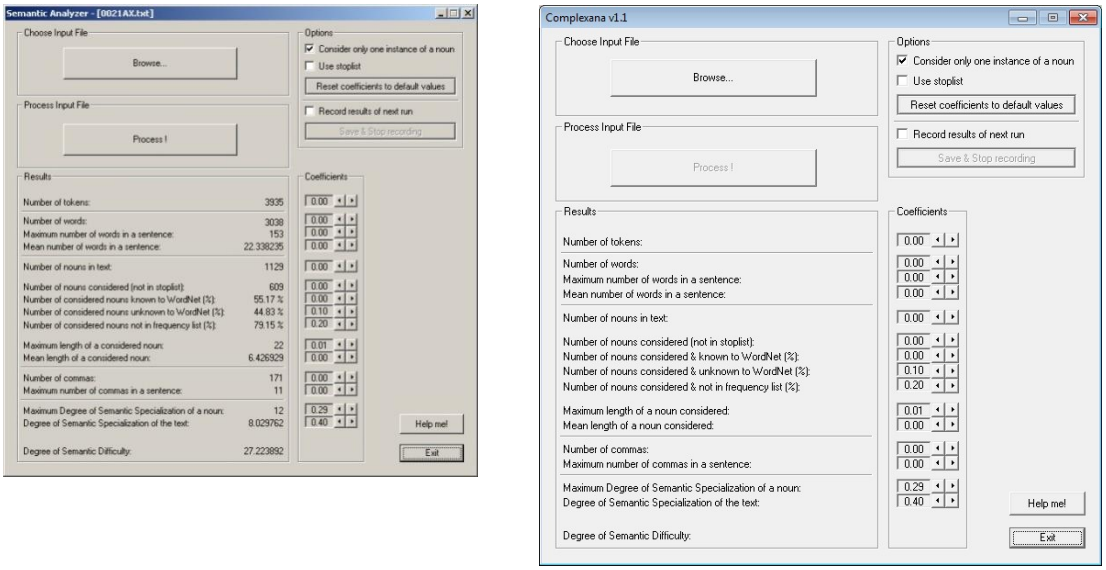

Figure 1: Two versions of ComplexAna, old and current

After a process of tagging all word classes, ComplexAna extracts all nominal items identified in the text and writes them to a separate file. Finally, a single score is calculated that summarizes the lexico-semantic complexity of the text as a dimensionless number. This score can only be compared to the scores obtained from other texts. We therefore define lexico-semantic complexity as a score that can only be considered relative and not absolute, as they make sense only in comparison.

Table 4: CUJOE word count in types and tokens and type/token ratio, see Haase, 2019

	tokens	types	TTR
all	488,417	24,665	0.0505
BA	297,958	18,978	0.063694
MA	190,459	13,070	0.068624

Further, the word number in the BA section is larger because it contains a larger number of bachelor theses. However, "as expected, the lexical variability is slightly higher for the high-proficiency learners when we compare the TTR (type/toke ratio) and it obviously decreases with larger corpus size, i.e. when we add both sections ('all') as more text simply (and lexically) means, 'more of the same'." (Haase, 2019).

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 8, Issue.1, Jan-Feb 2025, p no. 299-308

Table 5: Corpus text overview, see Haase, 2019

	mean length	minimum	maximum
all	14,429	3,609	49,665
BA	12,035	3,609	24,822
MA	20,414	8,251	49,665

Table 5 shows a survey of the single texts. Obviously, the MA theses are much longer by a large percentage (ca. 40%). Further, the distribution is not evn, the longest and the shortest text lie far apart in size.

4. Lexico-semantic complexity of CUJOE

As an additional recommendation, it is proposed to integrate tools like ComplexAna in teaching, as discussed in section 1, for learners to test their own texts with varying degrees of complexity. This enables students as novice practitioners to enhance the lexical profile of their texts and adjust style elements of their own academic output.

Table 6: Scores of lexico-semantic complexities in two sections of CUJOE

corpus file BA level	score	corpus file MA level	score
CUJOE001	22.83	CUJOE101	26.83
CUJOE002	24.09	CUJOE102	22.77
CUJOE003	24.74	CUJOE103	22.69
CUJOE004	23.65	CUJOE104	25.80
CUJOE005	22.68	CUJOE105	24.20
CUJOE006	23.28	CUJOE106	20.57
CUJOE007	22.38	CUJOE107	23.10
CUJOE008	24.36	CUJOE108	19.78
CUJOE009	26.11	CUJOE109	23.44
CUJOE010	24.04	CUJOE110	22.27
mean	**23.82**	**mean**	**23.14**
median	**22.93**	**median**	**23.84**

In Table 6, the scores exemplify the range of student writing in their accumulated complexity scores. As indicated in Table 5, the spread of diversity is higher for the more proficient students (MA section) than for the beginners (BA section). Even though the differences are not significant, the average score for the beginners is even higher. Due to the more heterogenous nature of the MA texts, the median gives a better metric. And here, the more advanced students score higher (23.84 in comparison to

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 8, Issue.1, Jan-Feb 2025, p no. 299-308

22.93). A breakdown of the text with the highest score summarizes the individual variables in Table 7.

Table 7: Summary of variables and their values from ComplexAna

Description	coefficients	CUJOE101
Number of Tokens	0	26743
Number of Words	0	20927
Maximum number of words in a sentence	0	90
Mean number of words in a sentence	0	12.822917
Number of nouns in text	0	8236
Number of nouns considered (not in stoplist)	0	1993
Number of considered nouns known to WordNet (%)	0	60.91
Number of considered nouns unknown to WordNet (%)	0.1	39,09
Number of considered nouns not in frequency list (%)	0.2	72.25
Maximum length of a considered noun	0.01	20
Mean length of a considered noun	0	6.95434
Number of commas	0	1394
Maximum number of commas in a sentence	0	14
Maximum degree of Semantic Specialization of a noun	0.29	17
Degree of Semantic Specialization of the text	0.4	8.343493
Degree of Semantic Difficulty		26.826654

The score is mainly driven by the high lexical specialization of the items, the relatively high sentence length and the overall length of the text. The individual breakdown thus enables students to control and adjust their own texts virtually in real-time and as a teaching implication, in classes on academic writing, methods can be implemented to boost the scores.

5. CONCLUSION

Through the use of specialized corpora and tools like ComplexAna, students can improve their academic writing skills by analyzing and optimizing the lexico-semantic complexity of their texts. The corpus SPACE and its accompanying tool have been widely used in academic writing coursework, English for Academic Purposes, and varieties studies, enabling learners to develop the necessary skills to write acceptable texts at the university level and for publication. The integration of corpus data and tools in teaching has also made assessment and evaluation more transparent and objective. Additionally, ComplexAna provides a comparative aid for linguists to understand transformation processes between different semantic levels of academic writing. Overall, the corpus and tool have proven to be a valuable resource for learners and researchers alike, enhancing their writing skills.

ISSN 2581-5148

Vol. 8, Issue.1, Jan-Feb 2025, p no. 299-308

In conclusion, the analysis of texts using corpus linguistic tools such as ComplexAna has revealed that lexico-semantic complexity is not solely determined by the frequency of difficult words as classified according to a standard ontology such as Wordnet. In the aforementioned examples we identified the impact of low-frequency lexical items. The integration of other parameters such as sentence length and the number of subclauses has been crucial in balancing the score and creating an accurate measure of lexico-semantic complexity. We believe that the use of tools like ComplexAna can be valuable for learners to test their own texts for readability and style, and to improve their writing skills. Additionally, the comparison of texts from different academic disciplines shows that different levels of lexico-semantic complexity apply, depending on the discpline. Overall, the integration of corpus data and tools like ComplexAna in teaching and writing can lead to a better understanding of academic vocabulary and improved communication in various academic fields for beginners and professional practitioners.

REFERENCES

Aston, G., Bernardini, S., & Stewart, D. (2001). *Corpora and language learners*. John Benjamins.

Aston, G. (2002). The learner as corpus designer. In Kettemann, B., & Marko, G. (Eds.), *Teaching and learning by doing corpus analysis*. Rodopi, 9-25.

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL* 17 (1), 47-64.

Bondi, M., & Scott, M. (2021). Corpus and discourse approaches to genre analysis: A study of research article introductions. *Journal of English for Academic Purposes*, 49.

Boulton, A., & Cobb, T. (2021). Developing and evaluating an online academic vocabulary resource for L2 learners. *System*, 98.

Braun, S., Kohn, K., & Mukherjee, J. (Eds.). (2005). *Corpus technology and language pedagogy*. Peter Lang.

Chambers, A. (2005). Integrating corpus consultation in language studies. *Language Learning & Technology* 9 (2), 111-125.

Chambers, A. & Thompson, P. (Eds.). (2007). Special issue on Incorporating Corpora in Language Learning and Teaching. *ReCall* 19, 3.

Chen, Y., & Baker, P. (2021). Exploring the use of phraseological patterns in EAP writing among Chinese and British university students. *Journal of English for Academic Purposes*, 50.

Cheng, W., Greaves, C., & Warren, M. (2021). The effectiveness of corpus tools in second language vocabulary learning: A meta-analysis. *System*, 99.

Fawcett, R. P., & Kuo, I. C. (2021). How corpora are changing the face of language assessment. *Language Testing*, 38(2), 165-183.

Fellbaum, C. (1998). Wordnet. An electronic lexical database. MIT Press.

Granger, S. (Eds.). (1998). *Learner English on Computer*. Addison Wesley Longman.

Granger, S. (2002). A bird's eye view of learner corpus research. In Granger, S., Hung, J., & Petch-Tyson, S. (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching*. John Benjamins, 3-33.

Granger, S. (2004). Computer learner corpus research: Current state and future prospects. In Connor, U., & Upton, T. (Eds.), *Applied corpus linguistics: A multidimensional perspective*. Rodopi, 123-145.

ISSN 2581-5148

Vol. 8, Issue.1, Jan-Feb 2025, p no. 299-308

Granger, S., E. Dagneaux, F. Meunier, & M. Paquot. (Eds.). (2009). *International Corpus of Learner English*. Version 2. UCL Presses Universitaires de Louvain.

Haase, C. (2009). Pragmatics through corpora in cultures: An empirical comparison of academic writing. *Topics in linguistics* 4, 23-30.

Haase, C. (2013a). The science of science in SPACE. In Haase, C. (Ed.)., *English for Academic Purposes: Practical and theoretical approaches*. Cuvillier.

Haase, C. (2013b). Tools for identifying and teaching semantic complexity in academic writing In Tafazoli, D. (Ed.)., *Language & Technology: Computer Assisted Language Teaching*, Khate Sefid, 129-137.

Haase, C. (2014). Registers of Analogy: Popular Science and Academic Discourse. In Vogel, R. (Ed.)., *Communication Across Genres and Discourses*. Muni Press.

Haase, C. (2016). Quantifying Lexico-Semantic Complexity in Academic Writing with Complexana. *International Journal of Language and Linguistics* 3, 6.

Haase, C. (2019). CUJOE - A New Academic Learner Corpus of English. In Haase, C. & N. Orlova (Eds.)., *English Language Teaching through the Lens of Experience*. Cambridge Scholars, 129-134.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. John Benjamins.

Renouf, A. & Kehoe, A. (Eds.). (2006). *The Changing Face of Corpus Linguistics*. Rodopi.

Sinclair, J. (2004). *How to use corpora in language teaching*. John Benjamins.