22 - The language of ChatGPT

Evelina Miščin
evelinamiscin@yahoo.co.uk
RIT Croatia, Zagreb, Croatia

Abstract
This research deals with the language of Artificial Intelligence (AI) with a focus on distinguishing characteristic vocabulary that delineates discourse pertaining to AI. The central inquiry revolves around identifying lexical patterns specific to AI discourse, thereby aiding educators in discerning when a paper is written by AI. Employing SketchEngine, we compare corpora from students' writings and ChatGPT, on analogous topics. Our methodology involves linguistic analysis to pinpoint unique vocabulary prevalent in ChatGPT corpus that distinguishes it from student-written texts. Preliminary results exhibit a distinct lexicon associated with AI discussions, including technical terms, jargon, and specialized terminology. These findings suggest the potential for developing linguistic markers to detect AI-centric content. The significance of this research lies in facilitating educators' ability to differentiate AI-related discourse, thus enhancing pedagogical practices and scholarly inquiry in language education. By establishing the distinct language of AI, this study contributes to a nuanced understanding of AI discourse and its implications for language education.

Keywords: AI-generated text detection, Academic integrity, Linguistic analysis, ChatGPT vs human writing, Educational implications of AI

1. Introduction

AI has become more and more prevalent in our daily lives. In the same way, it is finding its place in education. It could be useful for educators, helping them in various ways – from suggesting classroom activities, generating ideas for lesson plans, creating quizzes. Educators can also use it to facilitate debates by asking thought-provoking questions, or for generating case studies that promote discussion and critical thinking.

Despite its advantages, AI, in this case ChatGPT, can be misused by students who might use it for generating entire essays or projects without proper citation which leads to a lack of original thought and critical thinking. There are still no tools that can discover the use of ChatGPT or they are quite unreliable. Therefore, this research aims in investigating the language of ChatGPT trying to establish if there are certain structures and words used by ChatGPT which can help educators in recognising if the paper was written by a person or AI.

2. Theoretical background

There have been several studies dealing with the similar topic. For example, Mindner et al. (2023) explore methods to detect whether a text has been written by a human or generated by artificial intelligence (AI), specifically focusing on ChatGPT. They conducted experiments to classify basic and advanced human-generated and AI-generated texts, as well as AI-rephrased texts. The study includes the creation of a new text corpus covering ten school topics where the following features were used for text classification: perplexity, semantic, error-based, readability, list lookup and AI feedback features. The study achieved F1 scores above 96% for basic AI and human-generated text detection and over 78% for AI-rephrased texts. The paper concludes that combining traditional and new features can significantly improve the detection of AI-generated content, outperforming current tools like GPTZero.

[22] To cite this proceeding paper: **Miscin, E. (2024).** The Language of ChatGPT. In D. K.-G. Chan et al. (Eds.), *Evolving trends in foreign language education: Past lessons, present reflections, future directions. Proceedings from the 10th CLaSIC 2024* (pp. 249–258). Centre for Language Studies, Faculty of Arts and Social Sciences, National University of Singapore. https://doi.org/10.5281/zenodo.14504852

Georgiou (2024) explores linguistic differences between human-written and AI-generated texts. His study analyses various phonological, morphological, syntactical, and lexical features in both types of text using Open Brain AI, a computational linguistic tool. Among other findings, it is interesting to mention that AI-generated texts included more difficult words and content words, whereas human-written texts favoured easier words and function words. The study also emphasises the benefit of tools like Open Brain AI in linguistic analysis and assessment, particularly useful in education and healthcare. It concludes that despite high linguistic competence of AI-generated texts, there are clear differences between AI and human writing.

The article by Dugan et al. (2023) explores how well humans can detect transitions between human-written and machine-generated text. The authors used the RoFT platform, a gamified system where participants try to identify machine-generated sentences in various genres, such as news articles, stories and recipes. Their research established that certain genres, like recipes, were easier to detect compared to news and stories. The paper concludes that detecting AI-generated text remains a challenging but essential task, and suggests that with better tools and training, humans can enhance their ability to differentiate between real and fake text.

Another article, written by Berber Sardinha (2023) compares AI-generated texts, specifically those produced by ChatGPT, with human-authored text. It uses a multidimensional analysis approach based on the linguistic dimensions established by Biber (1988). The study indicates that AI-generated content, although sometimes resembling human language, still fails to fully capture the complexity of human communication. The conclusion is that while AI-generated texts can mimic human writing to some extent, they still exhibit artificiality, particularly in conversational and narrative contexts, showing that current AI models are not yet fully capable of replacing human-authored texts in various registers.

Similar research was carried out by Amirjalili et al. (2024) where they compared AI-generated texts and human-written academic texts in the context of English literature. The researchers compared an essay written by a second-year English literature student with a similar essay generated by ChatGPT-4. They analysed assertiveness, self-identification, and authorial presence using the "Voice Intensity Rating Scale" (VIRS). The paper highlighted the current limitations of AI in replicating the complexity and authenticity of human academic writing. The study suggest that educators must be cautious in how these tools are integrated into academic contexts, particularly in upholding academic integrity and encouraging genuine authorship.

All these studies explore how AI-generated texts differ from human writing and how these differences can be detected and analysed using various linguistic and computational approaches. They stress the importance of advancing detection tools and caution against over-reliance on AI in contexts requiring genuine authorship.

3. Research

The research was carried out in the fall 2023/2024 with Writing Seminar students. 56 students participated. Their task was to write a 2500-word essay on any topic of their choice. However, only 20 best essays were chosen for the research. Here is the list of titles of the chosen essays:

1. Climate change and its effects on health
2. Constant stress affecting students' mental health
3. The relationship between mental health condition and creative expression
4. Impact of hunger on cognitive function and memory recall
5. NBA vs the rest of the world
6. Rage to redemption: analysing Kratos' character development
7. Technologies in modern cinematography
8. The role of motonautica in student life
9. Reinvesting money to gain financial freedom
10. Protecting the human body to live in space and other planets
11. Developing social skills through video games
12. Down syndrome
13. Loot-boxes as in-game monetization system and their effect on the gaming industry

14. Qatar after FIF World Cup 2022
15. Albanian Besa
16. Will artificial intelligence make humanity smarter or dumber in the future?
17. Teaching strategies and outcomes: Finland and Croatia compared
18. Killer whales and the damaging effects of men on marine life
19. Differences in prosciutto production in Istria and Dalmatia
20. Feline Affection: Unravelling the Cat-Human Bond

After that, ChatGPT was asked to write the same length essays on the same topics. The aim of the research was to answer the following questions:

- How do the vocabulary and linguistic structures in AI-generated content differ from those in student-written texts on similar topics?
- What specific technical terms, jargon, collocations or specialized terminology are prevalent in AI-generated discourse that can be used as markers to detect such content?
- What are the pedagogical implications of being able to distinguish AI-generated language from human-authored writing in an educational context?

4. Procedure

Twenty best students' essays on various topics of their choice were collected and put into one file. On the other hand, ChatGPT 4.0 was asked to write the 2500-word essays on the same topics as mentioned before, for compiling the second file. Students' corpus consisted of 47,209 words and ChatGPT corpus of 37,748 words.

5. Results and discussion

Both files were separately analysed by *SketchEngine*. Sketch Engine's automatic keyword and terms extraction tool was used to obtain a list of the most frequent keywords, collocations and concordances (Figure 1).

Figure 1. Sketch Engine's interface, showing, among other features, the option "Keywords: Terminology extraction" that was used to extract the most frequent terms.

Both files were separately analysed. Function words were neglected and only content words were used for concordance analysis. After this, the results were compared.

First, it had to be established whether there are any significant differences between the two files. SketchEngine also offers this option as it can be seen in Figure 2.

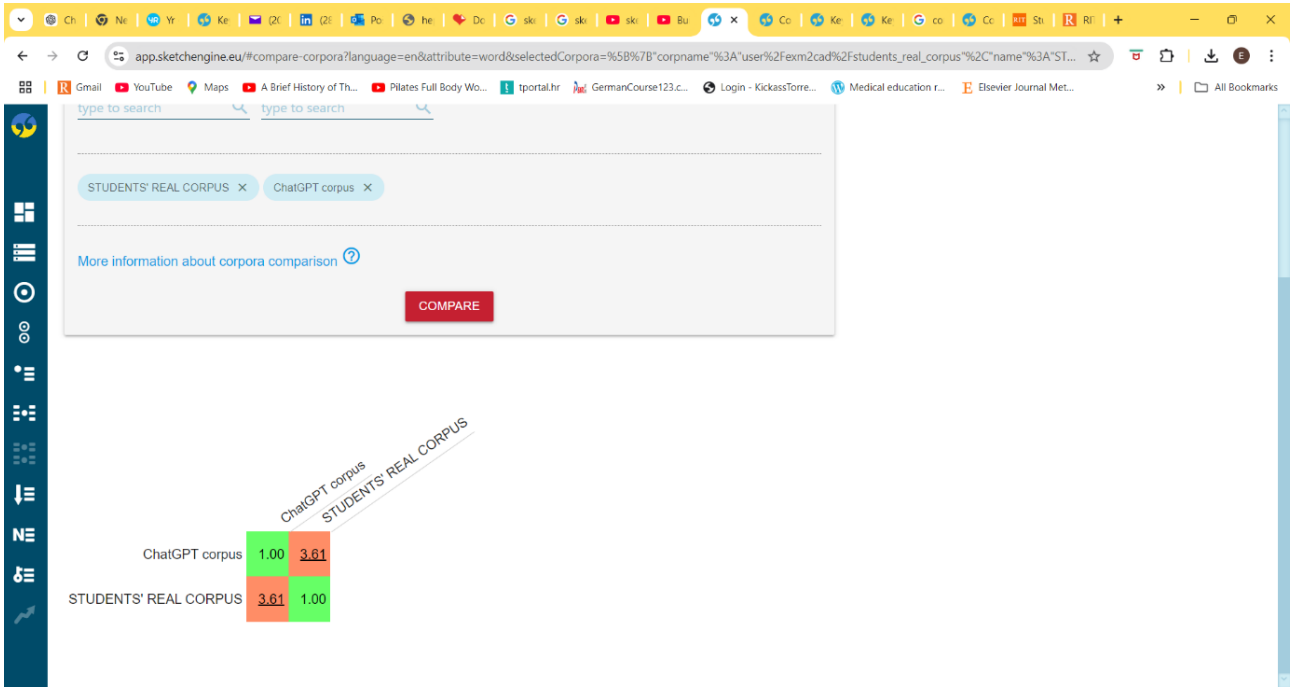


Figure 2. The difference between two corpora

As it can be seen in Figure 2, the difference between two corpora is 3.6. The value of 1 means identical corpora and the higher the score, the greater the difference between corpora. Therefore, the difference here is quite significant.

We shall start by analysing the most frequent words that appear in each corpus. Figure 3 shows 100 most frequent words in the students' corpus.

word (6,647 items | 54,547 total frequency)

	Word	Frequency	DOCF		Word	Frequency	DOCF		Word	Frequency	DOCF
1	,	2,969	1	35	people	126	1	69	important	80	1
2	the	2,591	1	36	but	124	1	70	research	78	1
3	.	2,223	1	37	was	124	1	71	cats	77	1
4	and	1,617	1	38	also	124	1	72	could	75	1
5	of	1,501	1	39	will	118	1	73	than	75	1
6	to	1,424	1	40	them	116	1	74	into	74	1
7	in	970	1	41	these	116	1	75	who	73	1
8	a	968	1	42	:	115	1	76	time	72	1
9	is	806	1	43	change	114	1	77	health	72	1
10	that	733	1	44	other	109	1	78	games	70	1
11	with	466	1	45	at	108	1	79	human	70	1
12	it	428	1	46	climate	103	1	80	social	69	1
13	as	427	1	47	life	103	1	81	such	68	1
14	for	407	1	48	there	101	1	82	students	68	1
15	are	397	1	49	how	101	1	83	about	67	1
16	)	382	1	50	all	101	1	84	education	67	1
17	(	382	1	51	mental	98	1	85	besa	67	1
18	this	358	1	52	skills	97	1	86	different	66	1
19	be	316	1	53	when	95	1	87	while	66	1
20	on	304	1	54	his	93	1	88	we	65	1
21	their	294	1	55	would	89	1	89	kratos	65	1
22	"	271	1	56	many	87	1	90	new	64	1
23	can	250	1	57	stress	87	1	91	financial	64	1
24	not	227	1	58	even	87	1	92	been	63	1
25	they	204	1	59	space	87	1	93	way	63	1
26	more	203	1	60	between	86	1	94	those	62	1
27	by	192	1	61	world	85	1	95	game	62	1
28	have	192	1	62	some	84	1	96	were	61	1
29	has	191	1	63	he	83	1	97	what	61	1
30	from	190	1	64	like	83	1	98	used	61	1
31	which	162	1	65	i	83	1	99	down	61	1
32	an	158	1	66	most	82	1	100	et	60	1
33	or	149	1	67	only	81	1				
34	one	140	1	68	its	81	1				

Figure 3. The most frequent words in the students' corpus

As it can been seen in the above Figure, even the punctuation marks appear in this frequency. As expected, the most frequent are function words and the first content word is 'people' which occurs in the 35. place.

For ChatGPT corpus, the result was a bit different and can be seen in Figure 4.

word (5,012 items | 42,704 total frequency)

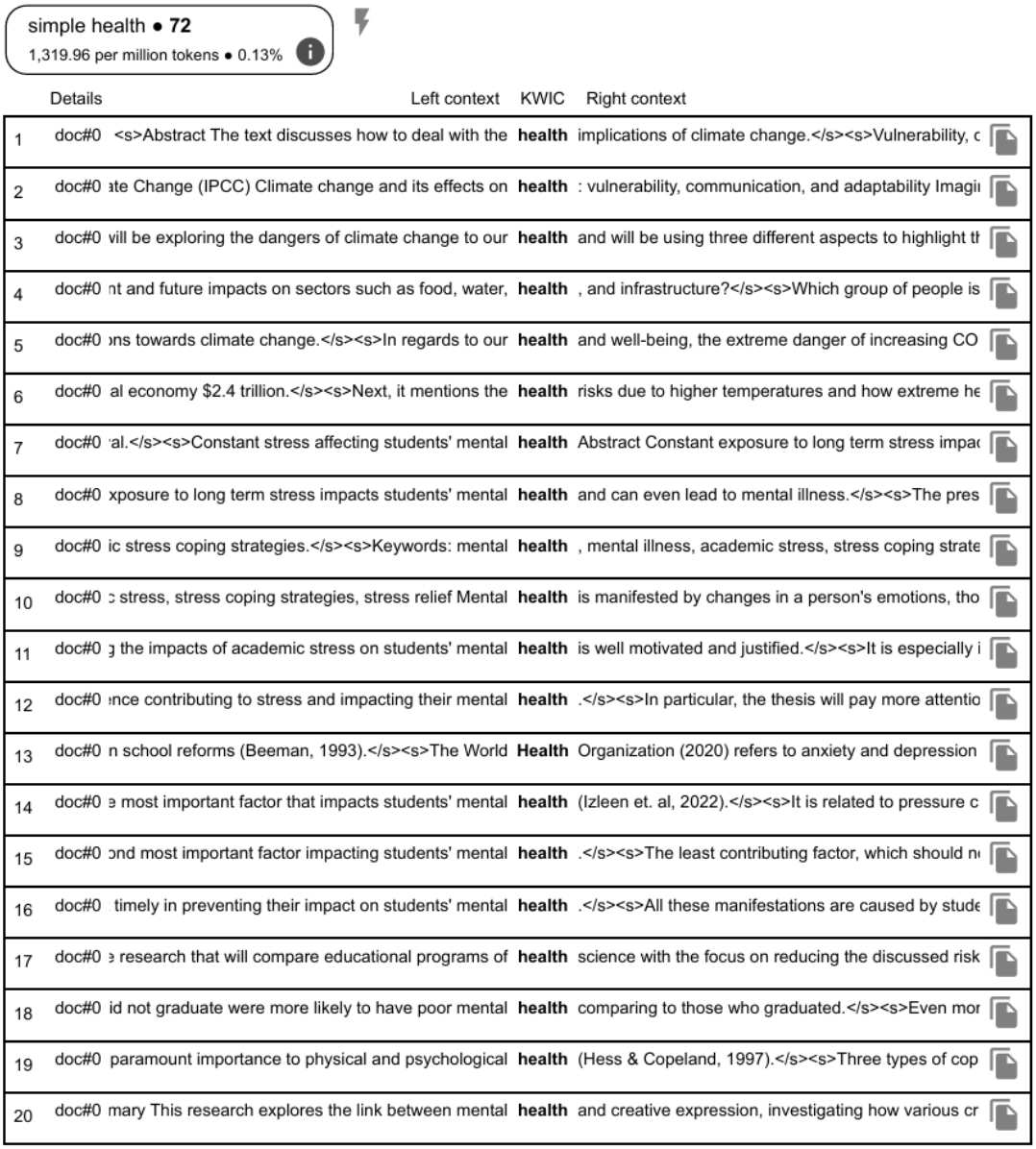
	Word	Frequency		Word	Frequency		Word	Frequency		Word	Frequency
1	,	2,497	26	from	131	51	individuals	76	76	down	61
2	and	2,421	27	has	122	52	killer	76	77	other	60
3	.	1,933	28	also	99	53	significant	75	78	potential	58
4	the	1,650	29	support	96	54	whales	74	79	while	57
5	of	1,178	30	or	94	55	more	74	80	syndrome	57
6	to	930	31	marine	90	56	cognitive	73	81	through	57
7	in	614	32	challenges	90	57	change	73	82	cats	57
8	a	610	33	games	90	58	international	70	83	crucial	57
9	can	509	34	its	90	59	often	70	84	lead	56
10	for	452	35	students	89	60	which	70	85	between	56
11	is	315	36	mental	87	61	systems	69	86	including	56
12	with	286	37	his	86	62	it	68	87	an	55
13	as	278	38	:	86	63	ai	67	88	creative	55
14	that	231	39	human	85	64	loot	67	89	effects	55
15	on	227	40	development	84	65	financial	67	90	strategies	54
16	this	222	41	provide	83	66	skills	66	91	role	51
17	are	206	42	climate	81	67	activities	65	92	hunger	51
18	health	201	43	(	79	68	boxes	65	93	global	50
19	their	196	44	impact	79	69	like	65	94	help	50
20	"	168	45	)	79	70	be	63	95	space	50
21	social	166	46	enhance	77	71	into	62	96	physical	50
22	by	166	47	education	77	72	essential	62	97	both	50
23	these	153	48	life	77	73	besa	62	98	cultural	49
24	such	150	49	players	77	74	world	61	99	promoting	49
25	have	132	50	stress	76	75	prosciutto	61	100	learning	49

Figure 4. The most frequent words in the ChatGPT corpus

The above figure shows that the function words are also the most frequent. However, the first content word appears a bit earlier than in the students' corpus and it is 'health', already in the 18. place. Both corpora share some frequent nouns – change, skills, stress, health, games, students, education, but they appear in different places. Words which only appear among the first 100 words of the students' corpus are: people, space, world, research, cats, time and besa, whereas those that appear only in the first 100 words of the ChatGPT corpus are: support, challenges, development, impact, life, players, individuals, whales and systems. From the semantic point of view, the students' corpus uses words more oriented towards personal, experiential or societal themes and the ChatGPT corpus uses words oriented towards scientific and social issues.

The next step was to compare concordances. The example is given for the word 'health' which appears as the most frequent in both corpora.

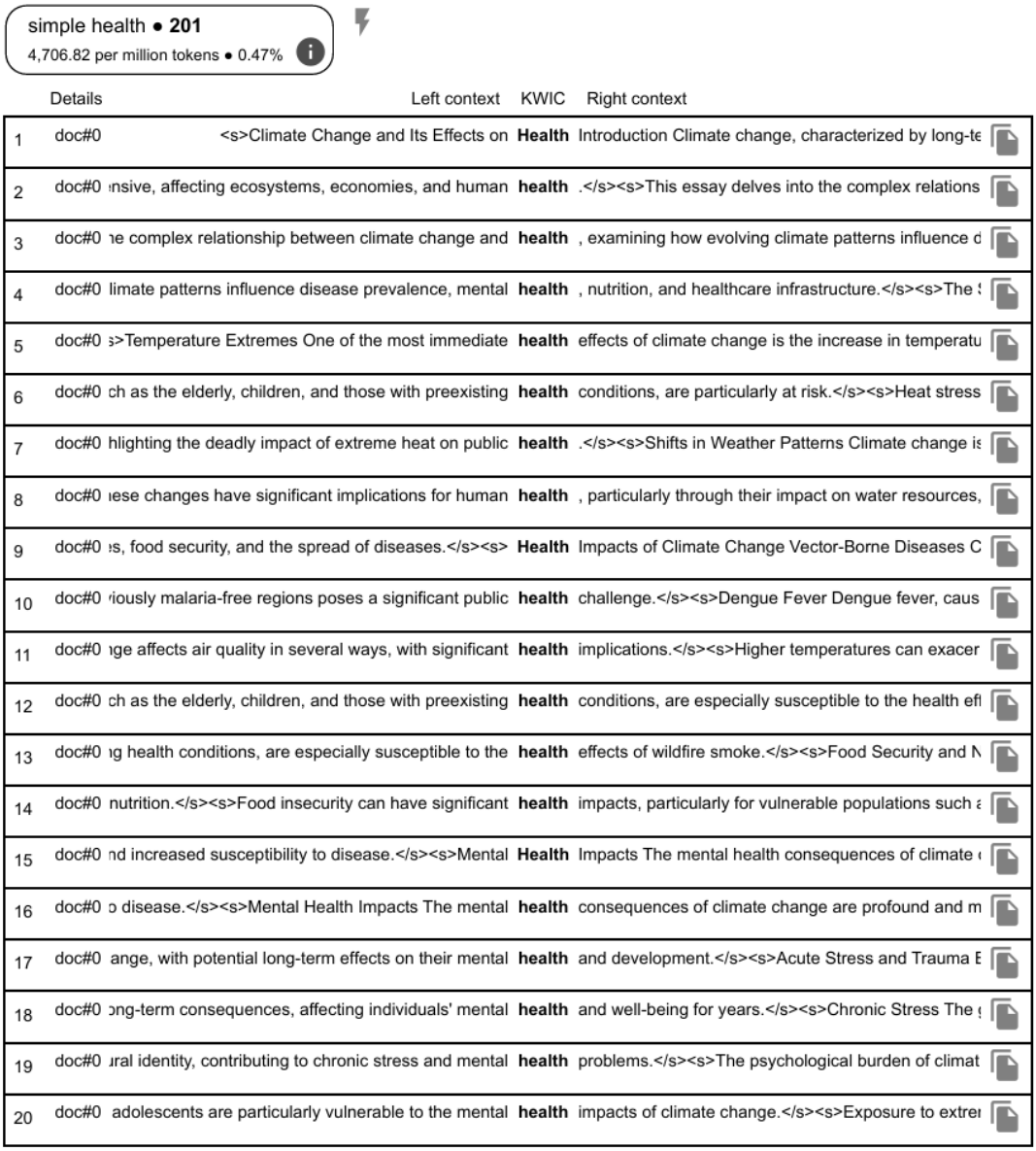
Figure 5 shows the concordances for 'health' in the students' corpus.



	Details	Left context	KWIC	Right context
1	doc#0	<s>Abstract The text discusses how to deal with the	**health**	implications of climate change.</s><s>Vulnerability, c
2	doc#0	ıte Change (IPCC) Climate change and its effects on	**health**	: vulnerability, communication, and adaptability Imagi
3	doc#0	vill be exploring the dangers of climate change to our	**health**	and will be using three different aspects to highlight th
4	doc#0	nt and future impacts on sectors such as food, water,	**health**	, and infrastructure?</s><s>Which group of people is
5	doc#0	ons towards climate change.</s><s>In regards to our	**health**	and well-being, the extreme danger of increasing CO
6	doc#0	al economy $2.4 trillion.</s><s>Next, it mentions the	**health**	risks due to higher temperatures and how extreme he
7	doc#0	al.</s><s>Constant stress affecting students' mental	**health**	Abstract Constant exposure to long term stress impac
8	doc#0	xposure to long term stress impacts students' mental	**health**	and can even lead to mental illness.</s><s>The pres
9	doc#0	ic stress coping strategies.</s><s>Keywords: mental	**health**	, mental illness, academic stress, stress coping strate
10	doc#0	c stress, stress coping strategies, stress relief Mental	**health**	is manifested by changes in a person's emotions, tho
11	doc#0	g the impacts of academic stress on students' mental	**health**	is well motivated and justified.</s><s>It is especially i
12	doc#0	nce contributing to stress and impacting their mental	**health**	.</s><s>In particular, the thesis will pay more attentio
13	doc#0	n school reforms (Beeman, 1993).</s><s>The World	**Health**	Organization (2020) refers to anxiety and depression
14	doc#0	e most important factor that impacts students' mental	**health**	(Izleen et. al, 2022).</s><s>It is related to pressure c
15	doc#0	ond most important factor impacting students' mental	**health**	.</s><s>The least contributing factor, which should n
16	doc#0	timely in preventing their impact on students' mental	**health**	.</s><s>All these manifestations are caused by stude
17	doc#0	e research that will compare educational programs of	**health**	science with the focus on reducing the discussed risk
18	doc#0	id not graduate were more likely to have poor mental	**health**	comparing to those who graduated.</s><s>Even mor
19	doc#0	paramount importance to physical and psychological	**health**	(Hess & Copeland, 1997).</s><s>Three types of cop
20	doc#0	mary This research explores the link between mental	**health**	and creative expression, investigating how various cr

Figure 5. Concordances for 'health' in the students' corpus

It can be seen that 'health' most frequently occurs in collocations like 'mental health', 'students' mental health', 'health science', 'health implications', 'psychological health'. Figure 6 shows concordances for 'health' in the ChatGPT corpus.



	Details	Left context	KWIC	Right context
1	doc#0	<s>Climate Change and Its Effects on	**Health**	Introduction Climate change, characterized by long-te
2	doc#0	ınsive, affecting ecosystems, economies, and human	**health**	.</s><s>This essay delves into the complex relations
3	doc#0	ıe complex relationship between climate change and	**health**	, examining how evolving climate patterns influence d
4	doc#0	limate patterns influence disease prevalence, mental	**health**	, nutrition, and healthcare infrastructure.</s><s>The
5	doc#0	s>Temperature Extremes One of the most immediate	**health**	effects of climate change is the increase in temperatu
6	doc#0	ch as the elderly, children, and those with preexisting	**health**	conditions, are particularly at risk.</s><s>Heat stress
7	doc#0	hlighting the deadly impact of extreme heat on public	**health**	.</s><s>Shifts in Weather Patterns Climate change is
8	doc#0	ıese changes have significant implications for human	**health**	, particularly through their impact on water resources,
9	doc#0	s, food security, and the spread of diseases.</s><s>	**Health**	Impacts of Climate Change Vector-Borne Diseases C
10	doc#0	ıiously malaria-free regions poses a significant public	**health**	challenge.</s><s>Dengue Fever Dengue fever, caus
11	doc#0	ıge affects air quality in several ways, with significant	**health**	implications.</s><s>Higher temperatures can exacer
12	doc#0	ch as the elderly, children, and those with preexisting	**health**	conditions, are especially susceptible to the health effl
13	doc#0	ıg health conditions, are especially susceptible to the	**health**	effects of wildfire smoke.</s><s>Food Security and N
14	doc#0	nutrition.</s><s>Food insecurity can have significant	**health**	impacts, particularly for vulnerable populations such a
15	doc#0	ıd increased susceptibility to disease.</s><s>Mental	**Health**	Impacts The mental health consequences of climate c
16	doc#0	ɔ disease.</s><s>Mental Health Impacts The mental	**health**	consequences of climate change are profound and m
17	doc#0	ange, with potential long-term effects on their mental	**health**	and development.</s><s>Acute Stress and Trauma E
18	doc#0	ɔng-term consequences, affecting individuals' mental	**health**	and well-being for years.</s><s>Chronic Stress The
19	doc#0	ıral identity, contributing to chronic stress and mental	**health**	problems.</s><s>The psychological burden of climat
20	doc#0	adolescents are particularly vulnerable to the mental	**health**	impacts of climate change.</s><s>Exposure to extrer

Figure 6. Concordance for the word 'health' in the ChatGPT corpus

The above figure shows that 'health' occurs in somewhat different collocations and phrases – for example 'health effects', 'health conditions', 'health impacts', 'health consequences', apart from the known one 'mental health'. There are also multi- word collocations like 'significant health impacts', 'preexisting health conditions', 'public health challenge'. Based on these two concordances, it can be concluded that in the ChatGPT corpus, 'health' is more associated with public health impacts, while the context in which 'health' appears in the students' corpus is more focused on individual well-being and educational setting, emphasising a personal perspective.

The last part of the research involved the analysis of collocations which occur in both corpora.

Figure 7 shows a hundred collocations in the students' corpus.

	Term		Term		Term
1	istrian prosciutto	35	students' mental health	69	psychological ownership
2	stress cope	36	creative outlet	70	gacha game
3	academic stress	37	deep space	71	developing social skill
4	down syndrome	38	blood feud	72	protein percentage
5	dalmatian prosciutto	39	multiplayer game	73	long-term depression
6	mental performance	40	don tran	74	albanian people
7	stress coping strategy	41	aspect of motonautica	75	mental well-being
8	creative expression	42	motoric ability	76	academic life
9	coping strategy	43	space ray	77	world of finance
10	fifa world cup	44	conscious sensation	78	gaming industry
11	santa monica studio	45	subjective happiness	79	young individual
12	sworn virgin	46	water percentage	80	star wars battlefront
13	cat behavior	47	human caregiver	81	visual effect
14	long-term potentiation	48	finnish education	82	emotional health
15	cognitive efficiency	49	memory retention	83	major sporting event
16	cooperative skill	50	mental health condition	84	western audience
17	problem-solving skill	51	sony computer entertainment	85	single father
18	space radiation	52	other aspect of life	86	type of job
19	emotional well-being	53	educational system	87	mental health challenge
20	creative activity	54	single-player game	88	augmented reality
21	killer whale	55	genshin impact	89	major sporting
22	space travel	56	sprained ankle	90	effect of climate change
23	monetization scheme	57	art therapy	91	chemical composition
24	synaptic weight	58	mental model	92	star player
25	warren buffet	59	loss of volition	93	effect of climate
26	emotional intelligence	60	entertainment-centric approach	94	professional support
27	emotional resilience	61	good cooperative skill	95	practical effect
28	cognitive performance	62	deep space ray	96	investment strategy
29	secure attachment	63	team-based video game	97	complete gacha
30	financial freedom	64	kind of loot-boxes	98	medicine in space
31	safety awareness	65	paid loot-boxe	99	cat-human bond
32	in-game monetization	66	aspect of life	100	source of academic stress
33	circadian rhythm	67	deep space travel		
34	social skill	68	attachment behavior		

Figure 7. Hundred collocations in the students' corpus

Apart from discipline-specific vocabulary, it can be seen that these collocations include abstract terms, like 'mental performance', 'emotional resilience', 'financial freedom', which denote concepts and qualities characteristic of academic writing. Multi-word terms like 'stress coping strategy', 'problem-solving skill', 'mental health challenge' reflect a linguistic economy, where multi-word terms serve to express complex ideas within a single phrase. Terms like 'psychological ownership' and 'attachment behaviour' demonstrate nominalisation, frequent in academic context to discuss abstract psychological and behavioural concepts. Phrases like 'aspect of life' and 'source of academic stress' serve pragmatic functions, such as generalizing or introducing complex ideas. These hedging expressions soften claims, making statements less absolute, which is a linguistic strategy often used in academic discourse to maintain objectivity.

Figure 8 shows a hundred collocations in the ChatGPT corpus

(items: **9,200**)

	Term		Term		Term
1	loot box	35	social dynamics	69	crucial role
2	killer whale	36	health impact	70	creative activity
3	dalmatian prosciutto	37	regional disparity	71	production technique
4	down syndrome	38	other planet	72	prey specie
5	marine ecosystem	39	social skill	73	educational outcome
6	creative expression	40	global cooperation	74	apex predator
7	financial freedom	41	mental health condition	75	pollution control
8	istrian prosciutto	42	cognitive impairment	76	marine mammal
9	international league	43	kratos' character	77	academic performance
10	prosciutto production	44	damaging effect of human activities	78	enabling filmmaker
11	extreme weather event	45	cat-human bond	79	principle of besa
12	cognitive function	46	in-game monetization	80	kratos' journey
13	feline affection	47	damaging effect	81	prosciutto industry
14	health of marine ecosystems	48	resilience of marine ecosystems	82	strong social bond
15	social interaction	49	direct human interaction	83	human activity
16	sustainable fishery	50	achieving financial freedom	84	overall well-being
17	conservation effort	51	availability of prey	85	long-term space
18	human companion	52	albanian society	86	social bond
19	impact of climate change	53	sustainable fisheries management	87	human interaction
20	impact of climate	54	effect of human activities	88	well-rounded individual
21	weather event	55	various aspect of life	89	adaptation strategy
22	habitat destruction	56	chronic hunger	90	real-time rendering
23	player experience	57	extra chromosome	91	education system
24	economic diversification	58	cultural impact	92	habitat restoration
25	health impact of climate	59	space exploration	93	skill development
26	chronic stress	60	teaching strategy	94	competitive balance
27	extreme weather	61	virtual production	95	fisheries management
28	reinvesting money	62	plastic debris	96	ethical consideration
29	modern cinematography	63	executive function	97	socio-economic background
30	health impact of climate change	64	talent development	98	fish population
31	marine life	65	gaming industry	99	environmental sustainability
32	geopolitical influence	66	space mission	100	early warning system
33	passive income	67	adequate nutrition		
34	student-centered learning	68	healthcare infrastructure		

Figure 8. Hundred collocations in the ChatGPT corpus

Collocations like ‘chronic stress’, ‘cognitive function’, ‘geopolitical influence’ are technical and scientific terms, indicating that the corpus provides factual information on many subjects. There are also very frequent abstract nouns like ’social dynamics,’ ‘economic diversification,’ ‘cultural impact.’ This supports corpus’s orientation towards the objective discussions. The formal tone is reinforced by terms like ‘conservation effort’ and ‘production technique,’ which describe systematic approaches to problem-solving in environmental and industrial contexts.

By comparing two corpora, it can be concluded that the tone of the ChatGPT corpus is more objective, while the student corpus is more subjective. The lexical choices in the ChatGPT corpus indicate high information density, commonly found in scientific literature. The student corpus uses terms that are less technical and more practical.

6. Conclusion

The analysis reveals distinctive linguistic characteristics between student-written and ChatGPT-generated content. The ChatGPT corpus demonstrates a more objective, technical tone, often using scientific and structured terms to discuss broad, factual topics, whereas the student corpus reflects a more subjective, experiential approach, with language oriented toward individual well-being, academic pressures, and personal engagement. The ChatGPT corpus exhibits high lexical density and formality, as seen in technical terms and structured collocations. In contrast, the student corpus

integrates conversational, accessible language and everyday vocabulary, making it suited for discussions directly relevant to personal experiences and educational themes.

Therefore, the answers to the questions posed in the introduction are as follows:

1. ChatGPT-generated content uses a high density of technical and scientific vocabulary with formal, objective language, focused on abstract, systematic topics. Student-written texts, in contrast, use more subjective language that reflects personal experiences.

2. Key terms and collocations in the ChatGPT corpus include 'cognitive function,' 'geopolitical influence,' 'health impacts,' and 'conservation effort,' which indicate a factual, technical focus.

3. Identifying AI-generated language can help maintain academic integrity, as students would be encouraged to submit original work and develop critical thinking and writing skills. For educators, having linguistic markers to detect AI text enables more accurate assessment of student comprehension and effort. This distinction can also guide curriculum development, supporting assignments that encourage authentic expression and discouraging over-reliance on AI-generated content.

The limitations of the study could be a limited set of topics selected by students. Another one could be overreliance on lexical markers and such a distinction could slowly disappear as AI are constantly evolving and may adopt a more human vocabulary. The analysis could be enriched by examining sentence structures, complexity, and use of passive versus active voice. Including examples of distinct syntactic patterns would strengthen the comparisons but not with the use of SketchEngine.

Future research could examine AI and human-authored texts across a broader range of genres and academic disciplines, from scientific reports to creative writing, to identify genre-specific linguistic markers. Further studies could incorporate syntactic and stylistic analyses, examining sentence length, complexity, and rhetorical devices. In cases of doubt about whether a student used AI, questions could be devised to verify the paper's authorship.

References

Amirjalili, F., Neysani, M., & Nikbakht, A. (2024). Exploring the boundaries of authorship: A comparative analysis of AI-generated text and human academic writing in English literature. *Frontiers in Education*, *9*, 1347421. https://doi.org/10.3389/feduc.2024.1347421

Berber Sardinha, T. (2023). AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, *4*, 100083. https://doi.org/10.1016/j.acorp.2023.100083

Dugan, L., Ippolito, D., Kirubarajan, A., Shi, S., & Callison-Burch, C. (2023). Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(11), 12763–12771. https://doi.org/10.1609/aaai.v37i11.26501

Georgiou, G. P. (2024). Differentiating between human-written and AI-generated texts using linguistic features automatically extracted from an online computational tool. *arXiv preprint arXiv:2407.03646.*

Mindner, L., Schlippe, T., & Schaaff, K. (2023). Classification of human- and AI-generated texts: Investigating features for ChatGPT. *arXiv*, *abs/2308.05341*.