

Alessia Battista

Sustainable Development Goal 4 across Institutional and Academic Discourses: An Analysis Integrating Corpus Linguistics and Artificial Intelligence

(doi: 10.7370/119425)

Textus (ISSN 1824-3967)

Fascicolo 3, settembre-dicembre 2025

Ente di afferenza:

()

Copyright © by Società editrice il Mulino, Bologna. Tutti i diritti sono riservati.

Per altre informazioni si veda <https://www.rivisteweb.it>

Licenza d'uso

Questo articolo è reso disponibile con licenza CC BY NC ND. Per altre informazioni si veda <https://www.rivisteweb.it/>

Sustainable Development Goal 4 across Institutional and Academic Discourses: An Analysis Integrating Corpus Linguistics and Artificial Intelligence

Alessia Battista

Abstract

This paper explores a selection of UN publications focusing on Sustainable Development Goal (SDG) number 4, which aims at fostering inclusive and equitable quality education and lifelong learning opportunities for all (United Nations n.d.-a), using a corpus-based approach. Two subcorpora will be analysed: open access scholarly papers and UN publications on SDG4 published between 2016 and 2024. The aim is to identify and compare the prevalent themes and discursive narratives about SDG4 across two genres, namely academic and institutional publications, as identified by a human researcher and a custom AI-powered tool combining corpus linguistics (Baker 2023) and Artificial Intelligence (Zappavigna 2023). Using the web-based platform Sketch Engine (Lexical Computing Ltd. n.d.), wordlists, keywords, and collocations will be explored. Additionally, this study will explore “discourse as representation” (Mahlberg 2014: 221) and cultural keywords (Bennett et al. 2005; Williams 2015), which are particularly relevant when dealing with issues of sociocultural significance, as is the case with SDGs. Then, the corpus will be examined by a custom GPT model, *Corpus Linguist* (Battista and OpenAI 2025), which has been created as an experiment aimed at understanding the potentialities (and limitations) of AI tools assisting human linguists. This analysis addresses the interdisciplinary dialogue between corpus linguistics and digital humanities by foregrounding the opportunities and methodological challenges of integrating AI and corpus linguistics (Smith et al. 2021) within the broader field of digital humanities.

Keywords: sustainable development goals, education, artificial intelligence, corpus linguistics.

1. Introduction

Digital humanities have transformed how scholars can study texts, language, and culture in the digital age. In linguistics research,

a number of computational methods and approaches, such as text mining and corpus analysis (Kirschenbaum 2012), have revolutionised English studies by enabling the exploration of large datasets and uncovering patterns that might be overlooked through manual analysis (Underwood 2019). This shift towards digital (often also quantitative) methods is termed the *digital turn* and reflects a trend of innovation in the humanities (Drucker 2021; Schwandt 2020).

Recent developments have shown how long-established methodologies and approaches, such as corpus linguistics, have been particularly receptive to this digital turn. Originated in the 1960s, corpus linguistics rapidly evolved as new technologies became available (Incelli 2025) focusing on the impact of artificial intelligence (AI, and by the 1990s new dedicated software enabled the empirical exploration of linguistic data (Sinclair 1991; Stubbs 2001). Technology is still influencing corpus linguistics and linguistics research in general, as the emergence of Artificial Intelligence (AI) has demonstrated, since the integration of AI-based tools in digital humanities offers unprecedented opportunities for analysis. For instance, Large Language Models (LLMs) can assist researchers in the precise detection of linguistic patterns and in the generation of insights from texts (Uchida 2024), underpinning a range of interactive tools solving complex tasks and engaging with users. Additionally, LLMs are highly scalable and adaptable, and present an invaluable advantage: they learn from unsupervised learning, which enables the model to improve its abilities simply by being used (Zhao et al. 2023). This principle is at the basis of Open AI's ChatGPT, which was launched in 2020 and has since been steadily growing in popularity and accuracy. At the moment of writing, the latest available version is ChatGPT 4.5, which can carry out complex tasks and generate and analyse images (OpenAI 2020).

Over the last couple of years there have been numerous studies exploring the potentialities and limitations of AI-based tools in linguistics, whose scalability, adaptability, and interactive potential have inspired scholars to create custom tools to perform complex linguistic tasks. However, current research also points to some limitations. On the one hand, scholars have argued that there might be some areas of communication which only humans can fully grasp (Bender and Koller 2020); on the other hand, researchers are trying

to apply AI specifically in these challenging areas, such as sarcasm detection (Battista 2025) and sentiment analysis (Wankhade et al. 2022). Results suggest that AI-powered tools might be limited in the analysis of multimodal and context-dependent texts (Castro et al. 2019; Zhao et al. 2023), and they might provide inaccurate and unreliable results in quantitative explorations (Curry et al. 2024), although they could still be useful to analyse large datasets from a corpus linguistics perspective (Zappavigna 2023). In particular, Curry et al. (2024) critically evaluated ChatGPT's ability to perform corpus-based tasks, such as keyword and concordance analysis, by comparing each individual task to that performed by humans, suggesting that AI has potential but also important contextual limitations and limited interpretative capabilities.

Starting from these considerations, and expanding the study by Curry et al. (2024), the present paper contributes to the ongoing discussions on AI and its possible intersection with digital humanities by presenting a custom AI-based tool, *Corpus Linguist* (Battista and OpenAI 2025), which has been developed as an experimental interface to explore the applicability of user-friendly, AI-based tools in corpus research. The aim is to assess its potential and limitations by systematically comparing its output with that of a human researcher using a 'traditional' corpus linguistics tool, namely Sketch Engine (Kilgarriff et al. 2014). To ensure the comparability and representativeness of the study, both the human researcher and the AI tool have focused independently on the same dataset and addressed the same research question. More specifically, the two analyses aim at identifying the prevalent themes and discursive narratives associated with Sustainable Development Goal 4 (SDG4), which promotes inclusive and equitable quality education for all. The topic has been chosen for its impact on policymaking and for the ideological and cultural narratives it may produce. It has been explored by scholars from multiple perspectives, highlighting the presence of important shortcomings in the fulfilment of SDG4, thus encouraging further reflections on the topic (for some recent reviews on the issue, see Edwards Jr et al. 2024; Hanemann 2019; Lafont-Torio et al. 2024; Sorooshian 2024) extent, and challenges to progress towards SDG 4: Quality Education for All at the mid-point of the 2030 campaign. Problematic paradigms, and potential pathways towards achieving Sustainable Development Goal 4. With

contributions from leading scholars and practitioners working in the areas of global governance, international development education, and comparative education, this special issue reflects on how far the world has come, provides clarity on what the fundamental obstacles to progress have been, and offers suggestions for ways forward, in addition to raising issues and posing (at times, uncomfortable). The corpus under scrutiny comprises a selection of open-access scholarly papers and publications by the United Nations published between 2016 and 2024 to explore how two discursive communities (i.e. academic and international) construct and frame the same SDG. Additionally, the two studies rely on corpus linguistics as a method (Baker 2023), and also consider the concepts of ‘cultural keywords’ (Bennett et al. 2005; Williams 2015), which highlight ideologically loaded and socially relevant terms, and “discourse as representation” (Mahlberg 2014: 221), which emphasises the connection between language and social reality. Therefore, the comparative approach that this study proposes seeks not only to analyse the representation of SDG₄ across genres, but most importantly to compare the analyses produced by a human researcher and an AI-based model.

2. Methods

As outlined in the introduction, the primary goal of this study is to compare two different corpus linguistics analyses carried out by a human researcher and by a custom GPT model. The aim is to assess whether and how the two analyses differ in terms of methodological approach and interpretive output. To make the analysis comparable, the same dataset was used, and the same research question was addressed, which aims at exploring the discursive and linguistic narratives around SDG₄.

Sustainable Development Goals were set by the United Nations in 2015 as part of the 2030 Agenda for Sustainable Development. They comprise 17 common goals to achieve to promote prosperity, protect the planet, and ensure peace and well-being (United Nations n.d.-b). One of them, goal number 4, aims to “ensure inclusive and equitable quality education and promote lifelong learning opportunities for all” (United Nations n.d.-a), and has been explored both by the United Nations (UN) and by academia. Therefore, an *ad hoc* corpus (*SDG₄_cor*) has been created, further

broken down into two subcorpora to enable the comparison across the two genres (table 1): the *UN_subcorpus* includes 10 official UN documents published between 2016 and 2024 focusing specifically on SDG4, while the *scholar_subcorpus* comprises the 10 most frequently cited open access research articles (according to Semantic Scholar (n.d.), which was selected due to its focus on high-impact and high quality research) published between 2016 and 2024 on SDG4. For comparative purposes, two reference corpora were selected among those available on Sketch Engine, namely the *Cambridge Academic English* (CAE) and the *United Nations Parallel Corpus – English* (UNPC).

TABLE 1
Details about corpora

Corpus	Subcorpus	Tokens	Types
SDG4_cor	UN_subcorpus	47,429	35,401
	Scholar_subcorpus	97,330	72,647
CAE	–	3,738,308	3,163,648
UNPC	–	798,630,043	664,924,245

All *SDG4_cor* texts were originally downloaded as PDF files, converted to plain text files, and cleaned of irrelevant metadata. While they could be analysed on Sketch Engine by the human researcher, the GPT model was unable to process the full corpus due to input length constraints and formatting complexity, which implied the need for the texts to be segmented into smaller files with no more than 10,000 lines and uploaded on the GPT model as a zipped folder. This procedure enabled *Corpus Linguist* to effectively access the data. The two subcorpora were analysed both separately and comparatively, both by the author (who will henceforth be referred to as ‘human researcher’) and by the GPT model.

Considering the ideological, social, cultural, and political relevance of the topic, which also entails issues of representation and inclusion, since it directly interconnects education with other dimensions – such as age, disability, gender – the methodological approach is informed by two key concepts. First, discourse patterns are regarded as reflective of the representation of social issues and

values in language, following Mahlberg's (2014) theory of discourse as representation. In particular, the present paper explores how discursive patterns may construct, reinforce, or even obscure social issues; indeed, the identification of these patterns was interpreted through a critical lens and with attention to their cultural and social implications in relation to issues of (in)visibility and hegemonic framing. Second, this study seeks to identify the presence of potential cultural keywords, which have been theorised as words that carry crucial cultural and ideological significance and evolve alongside social contexts (Bennett et al. 2005; Williams 2015), thus crystallising and reflecting fundamental societal debates. Indeed, because of the social significance of both discourse as representation and cultural keywords, this study hypothesises that such elements might be challenging for an AI-based model to identify due to the need for in-depth interpretation of language, which provides an additional layer of complexity.

TABLE 2
Sample items from wordlists and keyword lists and their assigned categories

Category	Five random items
UN / SDGs	development, sustainability, quality, agenda, goal
Education	education, learn, digital, study, school
Objects of learning / teaching	skill, literacy, science, competence, literature
Social actors	student, child, teacher, adult, youth
Actions / policies	ensure, policy, action, commitment, partnership
Research	show, result, research, model, analysis
Context	country, global, national, world, regional
Other	be, have, use, include, open

To avoid potential bias, the human-led analysis was carried out before the GPT analysis. Sketch Engine (Kilgarriff et al. 2014) was chosen, as one of the most sophisticated and established tools for corpus linguistics. Wordlists and keyword lists of verbs, nouns, and adjectives were generated (using the regex “V.*|N.*|J.*|”) for each subcorpus to investigate content words and identify prevalent themes. For the keyword analysis, both subcorpora were compared against both reference corpora to determine similarities and

differences within and across genres. Then, the keywords with a relative document frequency equal to or higher than 60 and with a keyness score equal to or higher than 5, and the top 50 items by relative frequency in the wordlists were semantically grouped. The semantic categories were not predetermined, but they emerged inductively from the inspection of the lists to identify dominant vocabulary as well as typical jargon regarding SDG₄ across genres (table 2).

Subsequently, potential cultural keywords were identified. First, the keyword lists produced by Sketch Engine were inspected by the researcher to retrieve those provided by Bennett et al. (2005) in their revised version of Williams' cultural keywords. Then, the corpus was manually inspected for additional potential items based on the features described by Williams (2015), namely that keywords are "significant, binding words in certain activities and their interpretation" and "significant, indicative words in certain forms of thought" whose uses "bound together certain ways of seeing culture and society" (Williams 2015: xxvii). Furthermore, keywords are considered as cultural if they meet the following criteria: carrying high ideological significance; reflecting societal debates; being contested, redefined, or reframed; having shifting meanings across contexts and/or time (Bennett et al. 2005; Williams 2015). Finally, the WordSketch tool was used to investigate the framing, collocations, and connotations of potential cultural keywords. While the two subcorpora are equivalent in terms of number of texts (10 each), they differ in size; therefore, in order to mitigate the effects of this imbalance and ensure a meaningful comparison, normalisation techniques were applied. Whenever possible, relative measures were considered (e.g. keyness score and relative (document) frequency), while percentages were preferred to raw counts in the presentation of frequency data.

Subsequently, *Corpus Linguist* was asked to perform a corpus linguistics study on the same research question and corpus. Developed by the author of this article, *Corpus Linguist* was conceived with the aim of simulating the work of a human corpus linguist. The model has been developed and trained for more than one year using the technical architecture created by OpenAI, which enables users to create a custom model without necessarily having programming skills, as they can simply rely on prompts written in a

natural language, such as English (Tabatabaian 2024; Wang and Jin 2023). However, although this process may seem intuitive and fairly accessible, it entails some major limitations due to the opacity of AI decision-making, often described as *black box*, according to which it is impossible for users to be fully aware of and control the internal reasonings of the AI (Kosinski 2024). *Corpus Linguist* has been trained to analyse and compare corpora, tokenising and lemmatising them, considering word frequencies, collocations, concordances, keywords. Additionally, the model can interpret data, since online searches are also enabled to access potentially useful social and cultural information, and it can generate visualisations, charts, graphs, and .xlsx files. The model has been trained to encourage user interaction by asking for clarifications and by allowing users to choose among multiple methods and statistical measures for collocation and keyword analyses. However, since this study aims at understanding the potentialities and limitations of the tool and at testing the model's autonomous decision making, input by the human user was minimal. In this case, *Corpus Linguist* decided to analyse wordlists, keywords, concordances, lexical fields, and modal verbs. Finally, the human and AI analyses were compared to identify similarities and differences.

Before illustrating the results, it should be mentioned that the design of this study was largely inspired by Curry et al. (2024). In three case studies, they tested ChatGPT's performance in corpus-based tasks, such as keyword grouping and concordance analysis, which have been compared to previously published human studies. Relying on Chat GPT's basic settings to mimic the queries of the average user, the study demonstrated that AI produces sensible yet superficial keyword grouping and poor performances for concordance analysis. Additionally, the scholars raised concerns about replicability, transparency, and data integrity in corpus-assisted discourse studies using ChatGPT. Building on Curry et al.'s study, the present paper employs a custom GPT model which has been trained by a linguist (the author of the present study) for linguists with the aim of performing more advanced corpus tasks than the free version of ChatGPT and of assessing whether targeted customisation can mitigate the limitations documented by Curry et al. (2024).

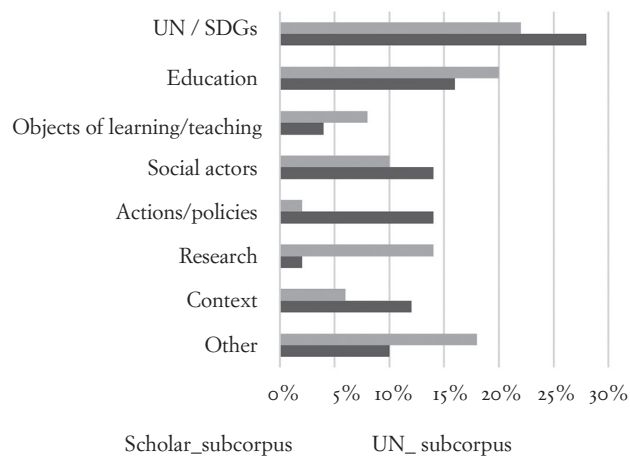
3. Results

This section presents the results of the analyses carried out by the human researcher and by *Corpus Linguist* separately. Then, they will be compared and commented in the discussion section of the paper.

3.1. Human analysis

The human-led analysis was carried out using Sketch Engine (Kilgarriff et al. 2014). Wordlists containing verbs, nouns, and adjectives were generated and semantically grouped by the researcher into conceptual fields to identify relevant themes. Figure 1 shows the percentage of words per category across the two subcorpora. It should be noted that the category “other” includes words that are too vague to belong to a specific group (most general-use verbs, such as *be*).

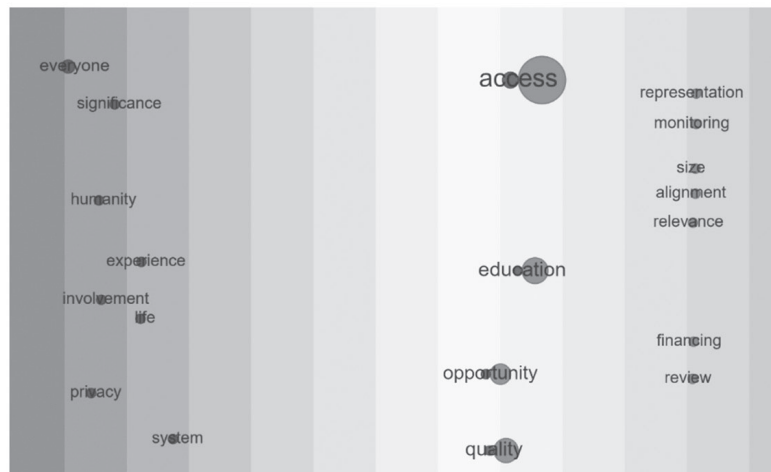
FIGURE 1
Selected wordlist items semantically grouped across subcorpora



As the graph shows, the thematic groups reveal that the two subcorpora have different orientations. In both subcorpora, the UN/SDGs and Education categories prevail, as could be expected

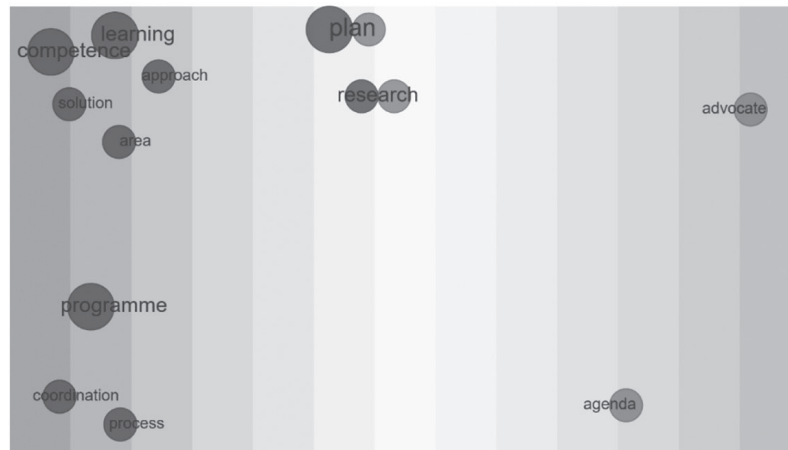
due to the topic. However, the *scholar_subcorpus* includes genre-specific words (category Other), which are not relevant for this study – such as *accessed* to introduce the date in which a referenced source was last accessed – followed by words describing measurements and theoretical frameworks (Research), as well as more detailed information on educational settings (Objects of learning/teaching and Social actors). Conversely, the *UN_subcorpus* focuses on actions and policies to pursue and details on their scope (Context, e.g. *country, global, regional*), as well as social actors who would benefit from them. Therefore, it can be concluded that academic publications focus on the human factor and have a scientific approach to SDG₄, whereas UN documents have an institutional and goal-oriented approach, as the presence of nouns denoting initiatives and verbs related to obligation (even moral) shows (figures 2 and 3).

FIGURE 2
WordSketch comparison of objects of ‘ensure’ in the *scholar_subcorpus* (left) and *UN_subcorpus* (right)



The wordlist analyses revealed the most frequent themes across the corpus. However, the analysis was broadened by considering keywords, indicating distinctive or even unexpected items for a genre.

FIGURE 3
WordSketch comparison of nouns modified by 'action' in the *scholar_subcorpus* (left) and *UN_subcorpus* (right)



Two reference corpora were selected as representative for the two genres under scrutiny, namely *Cambridge Academic English* (CAE) and *United Nations Parallel Corpus* (UNPC), and the two subcorpora were compared to both to unveil genre-internal and genre-external distinctiveness. The semantic categories are the same as the ones identified for the wordlists to make the results comparable, which are shown in table 3 as percentages (rounded up to the first decimal digit).

The *scholar-subcorpus* shows a thematic overlap between academic discourse and global education policy language when compared to the CAE, which indicates an intertextual alignment between academic and institutional language. However, when the UNPC is the reference corpus, the academic subcorpus shows greater specificity in content related to education and social actors, including keywords like *educator* and *teacher*, *learner* and *student* (category: social actors), but also *competence*, *innovation*, *lifelong (learning)*, *thinking* (category: Objects of learning/teaching), which are mostly absent in the *UN_subcorpus*. Therefore, not only does the *scholar_subcorpus* reference policy language, but it recontextualises it into pedagogical and epistemological frameworks.

TABLE 3
Selected keywords semantically grouped across subcorpora and compared to reference corpora

Semantic category	<i>Scholar subcorpus</i> vs CAE	<i>UN subcorpus</i> vs CAE	<i>Scholar subcorpus</i> vs UNPC	<i>UN subcorpus</i> vs UNPC
UN/SDGs	64.3%	34.6%	15.5%	31.7%
Education	21.4%	17.3%	20.7%	22%
Objects of learning/teaching	7.1%	5.8%	15.5%	9.6%
Social actors	7.1%	5.8%	6.9%	9.6%
Actions/ policies	0%	9.6%	5.2%	7.3%
Research	0%	3.8%	5.2%	0%
Context	0%	21.2%	5.2%	17.1%
Other	0%	1.9%	25.9%	2%

The *UN_subcorpus* shows a greater presence of keywords related to actions and policies, such as *partner* and *partnership*, *policy*, *ensure*, particularly when compared to the CAE, revealing the procedural aspect of institutional discourse, aiming at describing strategies, defining implementation actions, and showing accountability. It is also interesting to observe the shared keywords between the two subcorpora. Most of them relate to SDG4 and education in general due to the topic of the corpora, although some of them occur in different contexts. For example, in UN discourses *learning* typically appears in *lifelong learning opportunities for all*, a frequent phrase in SDG4 descriptions, while in the *scholar_subcorpus* it is often framed as a *transformative practice* and pre-modified by the adjective *critical*. The *UN_subcorpus* insists on commitment, highlighting the idea of inclusive and equitable quality education, whereas the *scholar_subcorpus* proposes a discursive expansion and occasional reinterpretation of SDG4-related concepts. Additionally, the keyword analysis makes it possible to observe some absences. Items related to gender, disability, marginalisation, which should be relevant when aiming at inclusive education, are only present in UN discourse and seem to be quite broad, such as *inequality*, *gender*, *gap*, *boy*, *girl*, as a random selection of concordance lines shows (table 4).

TABLE 4
Random concordance lines provided by Sketch Engine

Addressing	inequality	and ensuring inclusion in provision and in quality education outcomes requires deepening the understanding of teaching and learning in a given learning environment
Ensuring equity, inclusion and	gender	equality will require well-designed strategies
because the equity	gap	in education is exacerbated by the shortage and uneven distribution of professionally trained teachers
All girls and	boys	, women and men, should have equal opportunity to enjoy education
particularly for	girls	, women, vulnerable boys and youth, and other marginalized groups

This suggests that while UN discourse explicitly acknowledges their presence, in the academic subcorpus they might be irrelevant or embedded in issues of equity and access. Therefore, it is relevant to observe how inclusion is framed through collocation and concordance analysis.

Collocational analysis was carried out using Sketch Engine's WordSketch. The focus was on the shared keywords between the two subcorpora and on potential cultural keywords (Bennett et al. 2005; Williams 2015). While most shared keywords directly relate to SDG4, there are two broader items: *skill* and *program(me)*. The former is framed in UN discourse as a tool for empowerment, economic growth, and lifelong learning, with a focus on practical skills and employability, and the most frequent collocates include *competency*, *literacy*, *numeracy*, *writing*, *job*, *thinking*. Conversely, in the academic subcorpus skills are related to educational theories and/or empirical analyses, as can be predicted due to the nature of the texts. Some collocates include *practice*, *creativity*, *engagement*, *confidence*, as well as *OER* (Open Educational Resources). Similarly, the WordSketches of *program* (*UN_subcorpus*) and *programme* (*scholar_subcorpus*) reveal the same patterns, with *programs* frequently occurring with *education*, *training*, *curriculum*, *access*, *school*, but also with *implementation*, *framework*, *strategy* in

TABLE 5
Distribution of cultural keywords in the *SDG4_cor*

Candidate keyword	Present in Bennett <i>et al.</i> 2005 (y/n)?	Present in scholar_ subcorpus (y/n)?	Present in UN_ subcorpus (y/n)?	Collocates
<i>Education</i>	Y	Y	Y	<i>quality, secondary, equitable, basic, training, learning, opportunity</i>
<i>Quality</i>	N	Y	Y	<i>education, development, inclusion, affordable, enhancement</i>
<i>Equality/ Gender/ Inclusive</i>	Y (equality, gender); N (inclusive)	N	Y	<i>empowerment, diversity, equity, equitable, effective</i>
<i>Child/ Youth/ Girl/Boy</i>	Y (youth) N (child; girl; boy)	N	N	<i>all (girls and boys), adult</i>
<i>Citizenship</i>	N	N	Y	<i>global, empowerment, peace</i>
<i>Innovation</i>	N	Y	N	<i>creativity, technology</i>
<i>Learner</i>	N	Y	N	<i>motivation, value, skill, knowledge, educator</i>
<i>Student</i>	N	Y	Y	<i>school, friends, community, educator</i>
<i>Educator</i>	N	Y	N	<i>researcher, learner, empower, deliver, well-trained</i>
<i>Teacher</i>	N	Y	Y	<i>student, school, principal, shortage, training, support</i>

the *UN_subcorpus*, suggesting a policy-driven approach to create empowering programmes, and *programme* collocating with *study, outcomes, impact, model, evaluation* in the academic subcorpus,

thus discussing programmes as objects of academic scrutiny. It should be mentioned that some collocates are shared, including *well-being, resilience, confidence, (mental) health*, which clearly shows the impact of educational programmes on peoples' lives and mental health; however, in the UN documents this is framed as a goal to reach, while in academic discourse it is a variable to consider.

Subsequently, potential cultural keywords were identified based on Bennett et al. (2005) and on the keywords identified by Sketch Engine. Only four of the cultural keywords identified by Bennett et al. (2025) could be retrieved in the *SDG4_cor*, specifically *education, equality, gender, youth*. The other candidates emerged from the analysis of the corpus. A list of potential cultural keywords, occasionally presented as thematically grouped, their distribution and collocates is provided in table 5.

Based on WordSketches and concordances, table 3 shows the most frequent collocates as well as a final decision regarding the potential qualification of the candidate keywords as cultural. *Education* is reframed in this corpus as a right and a policy priority, while also needing rethinking and redefining following sustainability challenges, as exemplified in the concordance lines below:

- a) Education is a public good, of which the state is the duty bearer. Education is a shared societal endeavour, which implies an inclusive process of public policy formulation and implementation.
- b) Education is a fundamental human right and an enabling right. To fulfil this right, countries must ensure universal equal access to inclusive and equitable quality education and learning, which should be free and compulsory, leaving no one behind.
- c) Education is not only a fundamental human right but also a key driver in reducing poverty, improving health, fostering economic growth, and promoting peace and social cohesion.

The presence of *education* in both subcorpora and its variable collocational profile make it ideologically significant and occasionally contested; therefore, it fully qualifies as a cultural keyword. *Quality* is a more debatable item, since it is certainly an aim, but it is never clearly defined as shown in concordance lines.

- a) This is a serious problem, considering that teachers are important components in determining the quality of education itself.
- b) While providing access to learning for large numbers of learners, MOOCs have become instrumental in the transformation by institutions to more technologically enhanced learning, improving the quality of both online and blended learning. On the other hand, MOOCs have been criticised for the quality of their pedagogy, relying on traditional video lecture formats.
- c) SDG4 therefore pursues this unfinished education agenda, but also goes beyond, committing all countries to ensure equal opportunity in access to quality learning opportunities at all levels of education in a lifelong perspective.

Additionally, it strongly collocates with both technical and evaluative words, which make it ambiguous and highly context dependent. It is also continuously reframed, its meaning can change across context, it functions as a floating signifier (Lévi-Strauss 1987), and might thus be regarded as a cultural keyword. The keywords *equality/gender/inclusive* could be regarded as ideologically salient, and in the academic subcorpus some reflexive uses exist, but not enough to signal ideological and social reframing.

- a) Achieving gender equality requires a rights-based approach that ensures that girls and boys, women and men not only gain access to and complete education cycles, but are empowered equally in and through education.
- b) The UN 2030 agenda of Sustainable Development Goals (SDGs) envisions a future of inclusive equity, justice and prosperity within environmental limits, and places an important emphasis on education as stated in Goal 4.

As the above concordance lines show, their use is rather shallow and mostly formulaic, lacking specific definitions, and should rather be treated as a semantic field on rights, fairness, and access rather than as cultural keywords, although *equality* and *gender* have been classified as cultural by Bennett et al. (2005). The keywords *child/youth/girl/boy* were considered since they describe some strategic categories for political agenda. Below are some concordance lines:

- a) There is an urgent need for children, youth and adults to develop throughout life the flexible skills and competencies they need to live and work in a more secure, sustainable, interdependent, knowledge-based and technology-driven world.
- b) By 2030, ensure that all girls and boys have access to quality early childhood development, care and pre-primary education so that they are ready for primary education.

Nonetheless, they are rarely attributed agency or presented individually. Neither are they ideologically loaded nor contested in meaning; nor do they qualify as cultural keywords. *Citizenship* is a hotly contested topic in society and in academia; however, in the *SDG4_cor*, it evokes an abstract vision of the *ideal global citizen* which remains unclear as in the following concordance lines:

- a) The concept of global citizenship in GCED is critical, calling for proactive engagement for sustainable development, compared to the softer global citizenship component in ESD.
- b) Cross-sector approaches traversing education, science and technology, family, employment, industrial and economic development, migration and integration, citizenship, social welfare and public finance policies should be used.

Therefore, further studies are needed to understand whether it could qualify as a cultural keyword. Similarly, *innovation* is often related to progress and sustainability, but it remains quite vague, as exemplified in the excerpts below:

- a) Innovation and ICT must be harnessed to strengthen education systems, disseminate knowledge, provide access to information, promote quality and effective learning and deliver services more efficiently.
- b) What are the implications for national curricula, and what lessons have been learned from decades of experience of curriculum development about innovation that is pulled by users not pushed by providers, responds to effective demand, has a coherent epistemology, and contributes to national development priorities.

To conclude, the keywords describing social actors of education, namely *student* and *learner*, and *teacher* and *educator* have been explored. Sample concordance lines are shown in table 6 and discussed below.

TABLE 6
Concordance lines of cultural keywords for social actors of education in the *SDG4_cor*

Candidate keyword(s)	Sample concordance lines
<i>Learner</i>	From the perspective that we are advocating, teaching is based on the idea of making a concept clear and discernible to the learner. True inclusion involves creating environments that cater to diverse learners' needs, including investing in specialized training for teachers, providing accessible infrastructure, and fostering a culture of acceptance within schools.
<i>Student</i>	After the student selects the preferred class, the student will be directed to the class details page. Nearly 20 % of students had ideas to change how they learn – for example, “being able to experiment,” “choose the subjects they want to study,” or “study in nature.”
<i>Educator</i>	In thematic work with children, an educator can direct children's attention towards the environmental, sociocultural, and economic aspects of most topics in which children engage in early education. Teachers and educators, and their organizations, are crucial partners in their own right and should be engaged at all stages of policy-making, planning, implementation and monitoring.
<i>Teacher</i>	Although every city, region, and country have some kind of guidelines, curriculum, or framework for ECE that is known to staff within that context, teachers should be aware of the global agreements such as the SDGs. Such a conceptualisation also means that relationships, communication, and interaction have to be key notions of any ECE pedagogy for children's development of knowledge, skills, and attitudes, so that teachers can be a part of children's play world.

Learner and *student* are very strong candidates for cultural keywords, as in this corpus they reflect a shifting view of education from an active to a passive process (or vice versa, depending on the subcorpus). As a matter of fact, *learner* is framed as a passive subject to motivate, which implies the need to define agency in education,

while *student* is an active member of the school community, whose role may seem traditional but is nowadays under scrutiny for potential redefinition due to the new challenges in education. *Educator* is a cultural keyword since it reflects the emerging broader trend of lifelong learning opportunities, although more research is needed to fully understand how this role is framed, for instance as an expert, as an innovator, or as a facilitator. *Teacher*, instead, might be ideologically loaded and subject to reframing, since it is intended as a member of a broader community and as a profession, while also being the target of innovation in education due to some potential vulnerabilities emerging from collocates such as *shortage* and *support*. However, more research is needed to conclude if and how its meaning is shifting and thus to determine its status as a cultural keyword.

3.2. AI analysis

The research question addressed by the human researcher was also explored by *Corpus Linguist* (Battista and OpenAI 2025). The user initiated the conversation by providing the same dataset and research questions; then, since *Corpus Linguist* was designed as user-friendly, it asked some follow-up questions to set up the study, for instance asking whether the files should be treated individually or contrastively, and which methods to use. However, the model was allowed to use any measure or strategy it deemed as appropriate so as to limit human intervention. Appendix 1 shows the prompts and replies used to initiate the study.

Firstly, *Corpus Linguist* generated a lemma-based wordlist of each subcorpus and provided analytical insights. In the *UN_subcorpus*, a focus on institutional and programmatic terminology emerged, as well as a prescriptive, action-oriented approach; in the *scholar_subcorpus* a research-oriented approach was observed due to the presence of conceptual and normative frameworks, as well as academic vocabulary and citations. Frequent lemmas were grouped into thematic fields (*education_process*, *access_inclusion_equity*, *policy_governance*, *sustainability_values*, *implementation_action*). Their dispersion across texts revealed that policy, implementation and access, inclusion, and equity are minimal in scholarly documents, while in UN documents – which showed broader thematic coverage – sustainability is totally absent.

Secondly, the model generated a keyword list for each subcorpus using log-likelihood ratio, frequency filtering (items occurring at least 5 times to reduce statistical noise), and POS-tagging to identify cultural keywords (nouns) and to analyse representation (verbs). Noun keywords suggest that the academic corpus is concerned with the definition of educational models, whereas UN documents focus on policies. Verb keywords are scarce in the *scholar_subcorpus*, perhaps due to the preference for nominalisation in scholarly writing, while they are related to institutional authority and policymaking in the *UN_subcorpus*. Then, the model explored which keywords could qualify as cultural relying on KWIC (KeyWord In Context). Upon testing a set of candidates (*education, values, equity, justice, rights, inclusion, educate*), *Corpus Linguist* concluded that the most ideologically salient terms in the *SDG4_cor* are *values* and *educate*, since they are framed differently across the two subcorpora, and may thus be cultural. To further analyse the discursive framing of the two cultural keywords, KWIC lines were analysed. The human researcher did not impose any fixed parameter (e.g. 1R, 2R); the AI tool decided to sort KWIC lines based on the first word immediately following the keyword, since this makes it possible to identify the typical right-hand collocates and discursive frames. This revealed that in UN texts keywords are followed by assertive constructions, reflecting an institutional and prescriptive tone. Conversely, in academic texts they are followed by prepositional phrases implying a reflective approach and relational and conceptual associations related to the mission of (re)defining these concepts.

Thirdly, *Corpus Linguist* explored the use of modal verbs; however, it encountered some technical issues in analysing full sentences and could only extract information on *must*, observing that it occurs only once and only in the *UN_subcorpus*, commenting that this finding – albeit limited – reflected the policy-oriented nature of these texts.

To conclude, the model provided two tables before moving to its conclusions. Firstly, a concise, summarising table for its workflow (figure 4) was generated.

The table above shows the methods used by the model and the results they yielded. This may be the most effective way for *Corpus Linguist* to logically ground and gradually move to its results (figure 5).

FIGURE 4
Corpus Linguist's workflow

Objective	Method(s) used	Insight Produced
Identify prevalent themes	Wordlists, field mapping, dispersion	UN = broad coverage; Scholar = focused, thematic depth
Analyse discursive narratives	POS profiling, modal verbs, KWIC	UN = obligation + implementation; Scholar = analysis + representation
Explore cultural keywords	KWIC + right context, keyword lists	UN = implicit value signalling; Scholar = active cultural keyword discourse

FIGURE 5
Corpus Linguist's verdict

- 🔍 This study reveals how two discursive genres construct SDG4 through different representational strategies:
- 📖 **Scholarly texts** foreground SDG4 as a **conceptual and moral discourse**, invoking *equity, values, and sustainability* as sites of debate, and construct education as a **moral and ideological site**, foregrounding **critique, values, and epistemic inquiry**.
 - 🏛️ **UN texts** foreground SDG4 as an **institutional mandate**, using **assertive tone, modal language, and policy verbs** to construct education as both **target and tool**, and construct education as a **systematic object of governance**, emphasizing **measurable access, inclusion, and implementation**.
 - 🗣️ **Both discourses** align with SDG4, but do so in **structurally and ideologically distinct ways** — showing that **language encodes institutional priorities**, not merely reflects shared goals.
- 🗨️ Together, these findings support Mahlberg's view that discourse is not neutral description, but a means of representing the world — and Williams' insight that keywords are both cultural barometers and discursive battlegrounds.

As figure 4 shows, the AI model's final verdict is presented in a narrative style, without any reference to specific data, statistics or theory, and using some multimodal, not necessarily functional, cues such as bold typeface and emojis.

4. Discussion and Conclusion

Section 3 of this paper presented human-led and AI-led analyses of a corpus containing UN and academic publications on SDG4. The human researcher relied on a traditional corpus linguistics approach, using Sketch Engine, which provided quantitative insights based on wordlists, keyword lists, concordances and collocations, which led to

the identification of cultural keywords. *Corpus Linguist*, the custom AI model developed by the author, produced a more synthetic and narrative interpretation of the corpus, identifying prevalent themes and commenting keywords and collocates. Despite some technical limitations in the full retrieval of modal verbs, the cause of which is unknown and thus cannot be overcome, the model seemed able to use typical corpus linguistics measures and justify its choices; however, the transparency and replicability of the study is limited due to the black box (Kosinski 2024). This partly echoes Curry et al. (2024), who reported that the free version of ChatGPT could group keywords at surface level, but struggled with more in-depth levels of investigation.

Both analyses relied on similar methods and semantic approaches in the identification of prevalent themes and, within this dataset, they could both capture the core rhetoric on SDG4. Nonetheless, the human analysis provided deeper critical insights. Guided by the consideration of discourse as representation (Mahlberg 2014), it suggests a predominantly utilitarian view of education promoted in SDG4-related discourse (Brissett and Mitter 2017), and focused on collocational patterns across genres, suggesting that the *UN_subcorpus* has an institutional approach and a directive style, as opposed to the reflexive and critical approach of academic texts. Conversely, in line with other studies (Curry et al. 2024), the AI analysis is often less specific and context-sensitive than the human one; for instance, the keywords were generated by comparing the two subcorpora with each other and calculating log-likelihood scores for each lemma, without considering external reference corpora as they could not be provided. It should be clarified that the two analyses, while addressing the same research questions and investigating the same corpus, inevitably differ in scope. The human analysis exploits all the tools offered by Sketch Engine, including reference corpora, and could also foreground genre-specific distinctiveness. Conversely, the analysis by *Corpus Linguist* is constrained by the technical limitations imposed by the GPT's nature and could only compare texts and subcorpora. This crucial difference reflects this study's design, which does not intend to reproduce identical procedures, but aims at assessing the extent to which a custom AI model could (or could not) approximate corpus linguistics research. Therefore, while comparable in purpose, the two analyses

diverge methodologically: while the human researcher relied on established techniques and on statistics to comment the dataset and infer potential patterns, the GPT treated the corpus as a dataset to summarise, probably due to the fact that it could not process concordance lines and collocations as accurately as dedicated tools (e.g. Sketch Engine). Indeed, AI is known for adequately extracting and processing patterns (Zappavigna 2023); however, despite the freedom it was allowed in this study, it still needs human input to clearly define tasks, access pre-processed data and make sense of it, which might explain why the analysis seems limited. Additionally, upon manual (human) inspection of the corpus, GPT interpretations were not always accurate, indicating a need for human verification for interpretive steps, as was the case with the analysis of modal verbs. Finally, both analyses aimed at identifying potential cultural keywords; however, they adopted different approaches. The human researcher started from the list of keywords by Bennett et al. (2005) before identifying additional candidates based on Williams' criteria (2015); then, the collocations and concordances were manually inspected to validate the hypothesis. *Corpus Linguist* adopted a more superficial approach, as it identified potential cultural keywords by comparing how differently they were framed in the two subcorpora. While such an analysis can yield interesting insights, it may be limited in capturing the nuances needed for a lemma to fully qualify as a cultural keyword.

The findings of this study resonate with literature on SDG4 typically framing quality education as a human right and a catalyst for development (Vindigni 2024)2019. The identification of cultural keywords and the consideration of discourse as representation enabled ideological reflections and tied broader cultural assumptions to education, while also complementing discourse analysis by integrating social, political, and cultural values (Baker 2023). Academic discourse problematises and attempts to (re)define related concepts, as the focus on innovation and on multiple social actor shows, thus echoing the idea that cultural keywords can be revealing (Levisen and Waters 2017), whereas in UN texts their use remains more stable and formulaic, probably to adhere to genre conventions and seek political consensus. Within *SDG4_cor*, this contrast reflects differing ideological stances. Academic discourse positions itself as reflexive and ideologically sceptical, since scholars

are scrutinising initiatives related to SDG₄, while UN discourse is ideologically affirmative, as it presents the goal as necessary through its successes, leaving little room for questioning and debating basic concepts, including the definition of what could qualify as *quality* education. This is evident in the shallow and tokenistic treatment of gender, minorities, and disparities, mirroring existing research showing how the framing of SDG₄ typically universalises education and glosses over cultural and social divides (Brissett and Mitter 2017). This aligns with critiques suggesting that SDG₄ discourse reflects Western interests and assumptions (Arora-Jonsson 2023) and partially erases concepts like race, which has been regarded as an “absent presence” (Walker et al. 2023) that might be an obstacle towards the fulfilment of SDGs (Arora-Jonsson et al. 2019). These discursive absences are structurally embedded in institutional language seeking consensus and clarity. By omitting complex and politically charged framings of gender, race, marginalisation, UN discourse constructs a neutral narrative aligning with dominant ideologies and with the naturalisation of hegemonic assumptions (Fairclough 2001). This is visible in numerous expressions (see table 3) which invoke issues of inclusion and equality through vague and abstract language, without specifying involved social actors or structural causes. These discursive patterns frame inclusion and (in) equality as a measurable yet neutral goal, detached from structural injustice and power dynamics. Conversely, academic discourse hints at a problematisation of such concepts within broader ideological and cultural debates. The cultural implications are significant: relying on generic language, UN discourse marginalises experiences and identities, reinforcing a vision of inclusion that obscures structural issues and depoliticises inequities. In conclusion, while institutional discourse focuses on policy challenges, it also reproduces limited views of inclusion and equity, which might impact on how educational justice is understood and pursued.

Moving to the overarching research question regarding the possible integration of AI in corpus research, *Corpus Linguist*, despite still being under development and encountering some technical limitations, has proven capable of handling data and corpus tasks which would otherwise require more time and resources (Lin 2023) particularly its potential as a concordancer. As a corpus linguist and app developer, the author reflects on how ChatGPT’s

ease of use, efficiency, and popularity could challenge traditional concordancers, and explores ways in which ChatGPT could be used to generate concordances and frequency lists.”;”container-title”:”Applied Corpus Linguistics”,”DOI”:”10.1016/j.acorp.2023.100065”,”ISSN”:”2666-7991”,”issue”:”3”,”journalAbbreviation”:”Applied Corpus Linguistics”,”page”:”100065”,”source”:”ScienceDirect”,”title”:”ChatGPT: Friend or foe (to corpus linguists). Although it is not possible to provide an exact measurement, it can be estimated that the human researcher needed about four months to carry out the analysis – excluding the drafting of this paper – whereas the GPT only needed five hours, including prompt refinements.

Corpus Linguist has been trained to be able to interact with the user and carry out analyses independently; it can produce semantic summaries and detailed metareflections on its analytical processes, and it can provide methodological clarifications, making the analysis transparent and the model user-friendly, thus favouring broader participation in digital humanities (Uchida 2024; Zottola 2020). It can also spontaneously provide qualitative insights based on explicit patterns which are quickly generated (Incelli 2025), albeit limited and obscure due to the black box and in spite of custom training (for instance, it is unclear how some keywords were identified as potentially cultural). The methodological association between AI and corpus linguistics aligns with a growing body of research within the field of digital humanities. Not only could this be a new path for future studies (Zottola 2020) and an efficient tool for some types of annotations (Yu et al. 2024), but it is also naturally favoured by the existence of user-friendly tools for corpus analysis and widely accessible AI platforms. However, AI still poses numerous challenges due to the opacity of its internal processes and to the impossibility of replicating its output, as the same prompt might result in different responses (Berry 2022; Liao and Vaughan 2023). This has been tested for the study: while the (apparently) quantitative insights remained consistent, their interpretation changed from time to time, and it was necessary for the human user to ask it to revise its observations until no more changes were made, also based on prompt-sensitivity (McCallum and Mizumoto 2025). One last caveat should be made about the so-called researcher bias; although the use of a machine might seem helpful to overcome this, human users

have no access to the data used to train most LLMs and AI systems, which might be biased as well and might even perpetuate power dynamics in discourse (Gillings et al. 2024). To conclude, in spite of its limitations, AI can still be used effectively to identify main topics (Curry et al. 2024) and for pragmatic studies (Battista 2025; Yu et al. 2024). Nonetheless, current research in digital humanities is at a crucial point where AI's potential should be both embraced and critically evaluated to fully understand how it reshapes and contributes to research (Incelli 2025).

The present study has some limitations. Firstly, a highly specialised corpus was considered, which might not capture the full spectrum of discourses on SDG4. This choice was necessary because *Corpus Linguist* is still under development and cannot process larger datasets at this stage; therefore, to make the human and the AI analyses comparable, the dataset had to be the same. Secondly, the experimental nature of the GPT used might pose some limitations to its interpretative capacity. On the one hand, there are some technical constraints (e.g. lack of context sensitivity; black box; impossibility of replicating the study) which cannot be overcome; on the other hand, the GPT could become more sophisticated over time and use through RLHF (Reinforcement Learning from Human Feedback) (Dai et al. 2023). Finally, both analyses could have taken into consideration other tools, such as n-grams, and additional perspectives, for instance diachronic, to provide broader and more in-depth considerations. Nonetheless, due to time and space limitations, this study prioritised the concept of cultural keywords, which made it possible to make qualitative remarks starting from quantitative insights, thus highlighting the differences between human and AI research.

Future studies should build on the aforementioned limitations, by expanding the corpus to include a wider range of documents, such as reports, news, and social media discourse. It might also be interesting to carry out an AI-led analysis with human feedback to fully understand if and how GPTs can adequately and usefully support human researchers. It may also be possible to compare the analyses by *Corpus Linguist* and the free version of ChatGPT to more accurately determine the advantages of creating specialised models. Moreover, since this study focused on lexical patterns and keywords, other aspects of discourse might be integrated, such

as metaphors or argumentation, which could provide insightful results on discursive strategies and representation. By situating *Corpus Linguist* against the study by Curry et al. (2024), the present paper suggests that replicability and interpretive depth remain key challenges for AI, although targeted training can improve transparency and task articulation.

Overall, this contribution resonates with previous studies on SDG4 and its critiques, underscoring that discourses carry significant assumptions impacting on the portrayal of crucial issues and (under) represented identities. Methodologically, this study reinforces the connection between well-established corpus techniques and innovative AI-based tools which, in spite of extraordinary speed and adaptability, still entail some limitations in accuracy, consistency, and transparency. Therefore, it can be concluded that AI is a promising supplement, but not a substitute for human expertise, particularly for context-sensitive interpretations.

Appendix 1

Corpus Linguist 40 ▾

You are going to carry out a corpus linguistics analysis on SDG4. I will upload the corpus files and explain the aims of this study.

Great! Please go ahead and upload your corpus files. Once the files are uploaded, I'll also need the following details to set up the analysis properly:

1. How should the files be treated?
 - As a single combined corpus,
 - As individual documents, or
 - As separate sub-corpora for comparison?
2. What are your research questions?
 - Please list them clearly so I can align the analysis with your goals.
3. Which analyses do you want to run? (choose any that apply now or later)
 - Word frequency
 - Keyword extraction
 - Collocation (bigram/trigram)
 - Concordance (KWIC)
 - Named entity analysis
 - Phrase or n-gram extraction
 - Corpus comparison
 - Discourse-level or semantic analysis
4. Any specific statistical methods or filters? (e.g., MI for collocations, TF-IDF for keywords, entity filtering)
5. Do you want to exclude metadata? (e.g., titles, references, timestamps)
6. Should we follow a guided workflow to suggest the best methods based on your corpus and research questions?

Once you upload your files and provide these details, I'll take care of the rest. Ready when you are!

The corpus, SDG4 corpus, contains two subcorpora: Scholar and UN. They both include texts published between 2016 and 2024 on SDG4, but "scholar" includes the 10 most influential open access scholarly paper in (according to Semantic Scholar), while "un" contains 10 UN publications. Since the files are very dense and format-rich, I have decided to segment them. I will thus be uploading a zipped folder. It contains one folder per subcorpus; each subcorpus folder contains txt files (10 for cited and 5 for UN) with no more than 10000 lines each. The files have been numbered progressively, which should allow you to process each subcorpus and put the files in order. Before I provide the aims of this study, can you please confirm if the corpus design is clear so far?

Yes, your corpus design is clear and well-structured. Here's a quick summary to confirm:
[...]

Great! The aim of the study is to identify and compare the prevalent themes and discursive narratives about SDG4 across two genres, namely academic contributions and UN official publications, while also examining "discourse as representation" (Mahlberg 2014, 221) and cultural keywords (Williams 2014), which are particularly relevant when dealing with culturally and socially relevant issues, as is the case with SDGs. I want you to analyse each subcorpus individually and then also compare them with each other. Then, carry out the analysis step by step and keep me informed of everything you do. Feel free to use any measures you think could be useful. Do you have questions for me?

References

- ARORA-JONSSON, SEEMA, 2023, "The Sustainable Development Goals: A Universalist Promise for the Future", *Futures*, 146, 103087.
- ARORA-JONSSON, SEEMA, AGARWAL, SHRUTI, PIERCE COLFER, CAROL J., KEENE, STEPHANIE, KURIAN, PRIYA, LARSON, ANNE M., 2019, "SDG 5: Gender Equality – A Precondition for Sustainable Forestry", in P. Katila, C. J. Pierce Colfer, W. De Jong, G. Galloway, P. Pacheco and G. Winkel (eds.), *Sustainable Development Goals: Their Impacts on Forests and People*, C.U.P., Cambridge, pp. 146-177.
- BAKER, PAUL, 2023, *Using Corpora in Discourse Analysis*, Bloomsbury Academic, London.
- BATTISTA, ALESSIA, 2025, "Introducing sAIrcasm: A Sample Analysis of a Custom Artificial Intelligence for Linguistic and Discursive Sarcasm Recognition", *International Journal of Language Studies*, 19 (4), pp. 1-22.
- BATTISTA, ALESSIA and OPENAI, 2025, *Corpus Linguist*, accessible upon request, last accessed October 16, 2025.
- BENDER, EMILY M. and KOLLER, ALEXANDER, 2020, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data", in D. Jurafsky, J. Chai, N. Schlueter and J. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, pp. 5185-5198, <http://https://aclanthology.org/2020.acl-main.463>, last accessed June 27, 2024.
- BENNETT, TONY, GROSSBERG, LAWRENCE, MORRIS, MEAGHAN, WILLIAMS, RAYMOND, 2005, *New Keywords: A Revised Vocabulary of Culture and Society*, Blackwell, Malden.
- BERRY, DAVID M., 2022, "AI, Ethics, and Digital Humanities", in J. O'Sullivan (ed.), *The Bloomsbury Handbook to the Digital Humanities*, Bloomsbury Academic, London, pp. 445-457.
- BRISSETT, NIGEL and MITTER, RADHIKA, 2017, "For Function or Transformation? A Critical Discourse Analysis of Education under the Sustainable Development Goals", *Journal for Critical Education Policy Studies*, 15 (1), pp. 181-204.
- CASTRO, SANTIAGO, HAZARIKA, DEVAMANYU, PÉREZ-ROSAS, VERÓNICA, ZIMMERMANN, ROGER, MIHALCEA, RADA, PORIA, SOUJANYA, 2019, "Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper)", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4619-4629.
- CURRY, NIALL, BAKER, PAUL, BROOKES, GAVIN, 2024, "Generative AI for Corpus Approaches to Discourse Studies: A Critical Evaluation of ChatGPT", *Applied Corpus Linguistics*, pp. 100082.
- DAI, JOSEF, XUEHAI, PAN, SUN, RUIYANG, JI, JIANGMING, XU, XINBO, LIU, MICKEL, WANG, YIZHOU, YANG, YAODONG, 2023, "Safe RLHF: Safe

- Reinforcement Learning from Human Feedback”, *arXiv*, <https://arxiv.org/abs/2310.12773>, last accessed September 28, 2025.
- DRUCKER, JOHANNA, 2021, *The Digital Humanities Coursebook: An Introduction to Digital Methods for Research and Scholarship*, Routledge, Abingdon.
- EDWARDS JR, D. BRENT, ASADULLAH, M. NIAZ, WEBB, AMBER, 2024, “Critical Perspectives at the Mid-Point of Sustainable Development Goal 4: Quality Education for All—Progress, Persistent Gaps, Problematic Paradigms, and the Path to 2030”, *International Journal of Educational Development* 107, 103031.
- FAIRCLOUGH, NORMAN, 2001, *Language and Power*, Longman, London.
- GILLINGS, MATHEW, KOHN, TOBIAS, MAUTNER, GERLINDE, 2024, “The Rise of Large Language Models: Challenges for Critical Discourse Studies”, *Critical Discourse Studies*, pp. 1-17.
- HANEMANN, ULRIKE, 2019, “Examining the Application of the Lifelong Learning Principle to the Literacy Target in the Fourth Sustainable Development Goal (SDG 4)”, *International Review of Education*, 65 (2), pp. 251-275.
- INCELLI, ERSILIA, 2025, “Exploring the Future of Corpus Linguistics: Innovations in AI and Social Impact”, *International Journal of Mass Communication*, 3, pp. 1-10.
- KOSINSKI, MATTHEW, 2024, “What Is Black Box AI and How Does It Work?”, *IBM*, <http://https://www.ibm.com/think/topics/black-box-ai>, last accessed March 11, 2025.
- LAFONT-TORIO, JUAN, MARTÍN, JOSÉ MARÍA, SALINAS FERNÁNDEZ, JOSÉ ANTONIO, RIBEIRO SORIANO, DOMINGO, 2024, “Perceptions of Progress toward Achieving the Sustainable Development Goals: Insights from Cooperative Managers”, *Sustainable Technology and Entrepreneurship*, 3 (1), 100055.
- LEVISEN, CARSTEN and WATERS, SOPHIA, 2017, “An Invitation to Keyword Studies: Guidance for future research”, in C. Levisen and S. Waters (eds.), *Cultural Keywords in Discourse*, John Benjamins, Amsterdam, pp. 235-242.
- LÉVI-STRAUSS, CLAUDE, 1987, *Introduction to the Work of Marcel Mauss*, Routledge & Kegan Paul, London.
- KILGARRIFF, ADAM, BAISA, VÍT, BUŠTA, JAN, JAKUBÍČEK, MILOŠ, KOVÁŘ, VOJTĚCH, MICHELFÉIT, JAN, RYCHLÝ, PAVEL, SUCHOMEL, VÍT, 2024, “The Sketch Engine: ten years on”, *Lexicography*, 1, pp. 7-36.
- LIAO, Q. VERA and WORTMAN VAUGHAN, JENNIFER, 2023, “AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap”, *Harvard Data Science Review*, (5), <https://doi.org/10.48550/arXiv.2306.01941>, last accessed October 16, 2025.

- LIN, PHOEBE, 2023, "ChatGPT: Friend or Foe (to Corpus Linguists)?", *Applied Corpus Linguistics*, 100065.
- MAHLBERG, MICHAELA, 2014, "Corpus Linguistics and Discourse Analysis", in K. P. Schneider and A. Barron (eds.), *Pragmatics of Discourse*, De Gruyter, Berlin, pp. 215-238.
- MCCALLUM, LEE and MIZUMOTO, ATSUSHI, 2025, *Using LLMs for Corpus Linguistics Research: Promise, Disappointment and Progression*, Social Science Research Network, Rochester, <https://papers.ssrn.com/abstract=5224441>, last accessed May 29, 2025.
- OPENAI, 2020, *ChatGPT*, <https://chatgpt.com>, last accessed March 2, 2025.
- SCHWANDT, SILKE, 2020, *Digital Methods in the Humanities: Challenges, Ideas, Perspectives*, Bielefeld U.P., Bielefeld.
- SEMANTIC SCHOLAR, n.d., <https://www.semanticscholar.org/>, last accessed May 13, 2025.
- SINCLAIR, JOHN M., 1991, *Corpus, Concordance, Collocation*, O.U.P., Oxford.
- SMITH, THOMAS BRYAN, VACCA, RAFFAELE, MANTEGAZZA, LUCA, CAPUA, ILARIA, 2021, "Natural Language Processing and Network Analysis Provide Novel Insights on Policy and Scientific Discourse around Sustainable Development Goals", *Scientific Reports*, Nature Publishing Group, 22427.
- SOROOSHIAN, SHAHRYAR, 2024, "The Sustainable Development Goals of the United Nations: A Comparative Midterm Research Review", *Journal of Cleaner Production*, 453, 142272.
- STUBBS, MICHAEL, 2001, *Words and Phrases: Corpus Studies of Lexical Semantics*, Blackwell, Oxford.
- TABATABAIAN, MEHRZAD, 2024, *Prompt Engineering Using ChatGPT: Crafting Effective Interactions and Building GPT Apps*, Mercury Learning and Information, Boston.
- UCHIDA, SATORU, 2024, "Using Early LLMs for Corpus Linguistics: Examining ChatGPT's Potential and Limitations", *Applied Corpus Linguistics*, pp. 100089.
- UNITED NATIONS, n.d.-a, *SDG 4*, <https://sdgs.un.org/goals/goal4>, last accessed December 2, 2024.
- UNITED NATIONS, n.d.-b, *The 17 Goals*, <https://sdgs.un.org/goals#history>, last accessed May 13, 2025.
- VINDIGNI, GIOVANNI, 2024, "Overcoming Barriers to Inclusive and Equitable Education: A Systematic Review Towards Achieving Sustainable Development Goal 4 (SDG 4)", *European Journal of Arts, Humanities and Social Sciences*, 1 (5), pp. 3-47.
- WALKER, SHARON, TIKLY, LEON, STRONG, KRYSTAL, WALLACE, DERRON, SOUDIEN, CRAIN, 2023, "The Case for Educational Reparations: Addressing Racial Injustices in Sustainable Development Goal 4", *International Journal of Educational Development* 103, 102933.

- WANG, SHUYUE and JIN, PAN, 2023, "A Brief Summary of Prompting in Using GPT Models", *Qeios*, <https://www.qeios.com/read/IMZI2Q>, last accessed March 2, 2025.
- WANKHADE, MAYUR, RAO, ANNAVARAPU CHANDRA SEKHARA, KULKARNI, CHAITANYA, 2022, "A Survey on Sentiment Analysis Methods, Applications, and Challenges", *Artificial Intelligence Review*, 55 (7), pp. 5731-5780.
- WILLIAMS, RAYMOND, 2015, *Keywords: A Vocabulary of Culture and Society*, O.U.P., Oxford.
- YU, DANNI, LI, LUYANG, SU, HANG, FUOLI, MATTEO, 2024, "Assessing the Potential of LLM-Assisted Annotation for Corpus-Based Pragmatics and Discourse Analysis: The Case of Apology", *International Journal of Corpus Linguistics*, 29 (4), pp. 534-561.
- ZAPPAVIGNA, MICHELE, 2023, "Hack Your Corpus Analysis: How AI Can Assist Corpus Linguists Deal with Messy Social Media Data", *Applied Corpus Linguistics*, 3 (3), 100067.
- ZHAO, WAYNE XIN, ZHOU, KUN, LI, JUNYI, TANG, TIANYI, WANG, XIAOLEI, HOU, YUPENG, MIN, YINGQIAN, *et al.*, 2023, "A Survey of Large Language Models", *arXiv*, <http://arxiv.org/abs/2303.18223>, last accessed June 5, 2024.
- ZOTTOLA, ANGELA, 2020, "Corpus Linguistics and Digital Humanities. Intersecting Paths. A Case Study from Twitter," *América Crítica*, 4 (2), pp. 131-141.