Introduction

Recent years have witnessed education becoming an increasingly global phenomenon, in which students from all linguistic cultural, and ethnic backgrounds are invited to participate in conventional and distance learning activities (Douglas, 2016). With the advent of digital technology and the internet, cross-cultural dissemination of knowledge has increased to unprecedented levels. For studies of linguistics, Corpus Linguistics (CL) is among the most dynamically evolving domain, which encompasses the nuances of modern linguistics. Due to its vast breadth of knowledge, it influences all the areas of linguistics including methods of teaching of local and foreign languages, linguistic disciplines in cross cultural contexts, and complex translations (Nartay et al., 2019). It is based upon the development and use of a Corpora, which is its body of‘ real world' text. Corpus linguistics propounds that credible assessment of any language is feasible through the corpora collected in the field, that is, through the natural context of that language (realia) rather than experimental interference.

The concept of Corpus linguistics has been practiced in various eras in the form of large collection of literary writing. For example, early Arabic literary scholars used the locally developed corpus of Arabic language to comprehend the religious scriptures of the Quran (Zeroual et al., 2018). Similarly, lexicographers of the English language have been developing large samples of European languages to facilitate in accurately describing a word since the early 19th century (Adams, 2016). With the emergence of digital computers, the process of scribing has shifted to digital tools, and the internet exponentially expanded the volume of repository. The first computer-based corpus constituted of 1 million words. Whereas, generalised corpora today include hundreds of millions of words, which helps the non-speakers in gaining in-depth perspective about research and teaching of the foreign language. A significant contribution in this field occurred in the form of ‘Computational Analysis of Present-Day American English' by Henry Kučera and W. Nelson Francis in 1967 (Neslon, 2019). This work was constituted upon the Brown Corpus, which was a voluminous compilation of about one million American-English words of that period of time. Meanwhile, the authors demonstrated the use of corpus outside of traditional teaching and learning, to understand its applications in a diverse number of fields such as statistics, psychology, sociology, along with linguistic teaching.

Since the modern education system in the United Kingdom has become highly inclusive in terms of culture and ethnicities, the current research aims to investigate the impact of the characteristics of Corpus linguistics on the linguistic education of non-European students. Therefore, this research intends to realise the following objectives:

- To assess the impact of Corpus design and corpus sampling in Corpus linguistics on education attainment amongst non-European students in the UK
- To evaluate the role of presentiveness of non-European cultures in effective learning of non-European students in the UK

Literature Review and Hypotheses Development

A corpus is an expansive and principled collection of texts of a particular language collected and stored in electronic format (Reppen, 2019). Despite the absence of a standardised size or format of text collection, early corpus used to include up to 1 million words (Love et al., 2017). This standard was implicitly set by the limits of human capability as well as complications in data management faced during the decades of 1960s and 1970s. However, digital computers facilitated in exponentially increasing the standard size of corpora, while the internet enabled the corpora developing scholars in gaining quick access to an ever-increasing body of English vocabulary (Zerkina et al., 2017). Although certain well-known specialist corpora are far smaller than that, it is generally believed that bigger corpora are more beneficial for most corpus linguistics tasks. One other characteristic of contemporary corpora is that they are frequently made available to other researchers, generally for some predefined fee and occasionally for no cost (Yu et al., 2019). This is an important advance since it makes it possible for academics from across the world to access the same sets of data, encouraging more responsibility in data analysis and allowing for collaborative work and follow-up study from other scholars. Timmis (2016) asserted that teaching and learning process in various levels of schools have also facilitated substantially due to this expansion of corpora. In the UK, the field of Corpus linguistics has become a part of undergraduate, graduate, and doctorate level education, in which researchers learn to apply digital tools and techniques to analyse the textual data obtained from a corpus.

Since corpus are inherently expansive in terms of textual information, a leading characteristic that determines their effectiveness is their design. Design of the corpus influences the kinds of analyses that can be carried upon it, and it also affects the reliability of the results obtained from its analysis (Suarez et al., 2020). Hence, composition of a corpus should be in line with the research goals of the study carried upon it, and vice versa (Laviosa, 2021). For example, any corpus purposed towards exploration of lexical questions has to be considerably large in order to incorporate accurate representation of voluminous words along with their myriad of senses, meanings, and context-based usage. Whereas, for grammatical explorations, size is not a constraint because considerably fewer grammatical construction exists compared to lexical construction, and thus they recur more frequently (Rogers et al., 2021). Nonetheless, corpus design and its compilation need to reflect the issue being investigated through it. With this analysis, following hypothesis is developed which relates the design and compilation of corpus with its effectiveness on education for the non-European students.

H1: Corpus design and compilation has a significant impact on education of non-European students in the UK

Intensive utilisation of the internet technology has enabled linguistic researchers to access linguistic data in a depth hitherto underheard of. Extending beyond the dimension of number of texts, corpus developers are now capable of analysing numerous historical and contemporary themes, their historical evolution, their modern aspects, and their relation and placements in the overall body of language (Friginal, 2018). However, efficacy of such an expanded analysis is influenced by the breadth of sampling done by the researcher. Since a corpus is required to be principled, it is contingent that the words and the language are non-random, and are chosen due to specific characteristics (Rogers et al., 2021). A principled corpus is also important for creating large generalised corpora, because the user may use them to establish generalisations regarding their specific findings (Rogers et al., 2021).

The Brown Corpus, the LOB Corpus, the COCA, or the BNC are examples of generic corpora whose aim is to reflect language in its broadest sense and to offer a thorough grasp of the structure and function of language (Reppen, 2019). More often than not, general corpora are designed to be rather large. For instance, the BNC had 100 million words when it was first formed in the 1990s, and the COCA had 560 million words in 2019. (Rogers et al., 2021). Research of Mair (2015) has shown that one million words are sufficient to give reliable, generalizable solutions for some analyses, albeit not all research concerns, even if the one-million-word size of the early general corpora like Brown and LOB makes them look modest by today's standards. A general corpus is made to be balanced and contain linguistic samples from many different registers or genres, encompassing both fiction and non-fiction in all their variety.

Specialised corpora, such as the TOEFL-2000 Spoken and Written Academic Language Corpus, and the International Corpus of English (ICE), a corpus created for the study of national variants of English, may have both spoken and written components (Friginal et al., 2017). A specialised corpus, more frequently, concentrates on a specific spoken or written variant of language. Historical corpora, like the Helsinki Corpus, are among the specialised written corpora. Meanwhile, a new trend of corpus that is getting considerably importance amongst language teacher is the ‘Learner’s corpora’ (Gilquin and Granger, 2015). This kind of corpus comprises of written and spoken samples of language developed non-native speakers. One example of this corpus is the International Corpus of Learner English (ICLE). These non-native speakers are generally L2 level learners for whom, the language is their second language (Gilquin and Granger, 2015). Learner corpus has emerged as the prime example of the tremendous potential of non-native speakers in developing the breadth of knowledge about a particular foreign language. Meanwhile, avenues such as the Learners’ corpus also facilitate students in gaining both, the experience of contributing to a foreign language and the confidence for future research in that foreign language (Gilquin, 2022). Through the above analysis, it can be hypothesised that sampling of the corpus may have a substantial influence on the education of non-European students in the UK.

H2: Corpus sampling has a significant impact on the education of non-European students in the UK

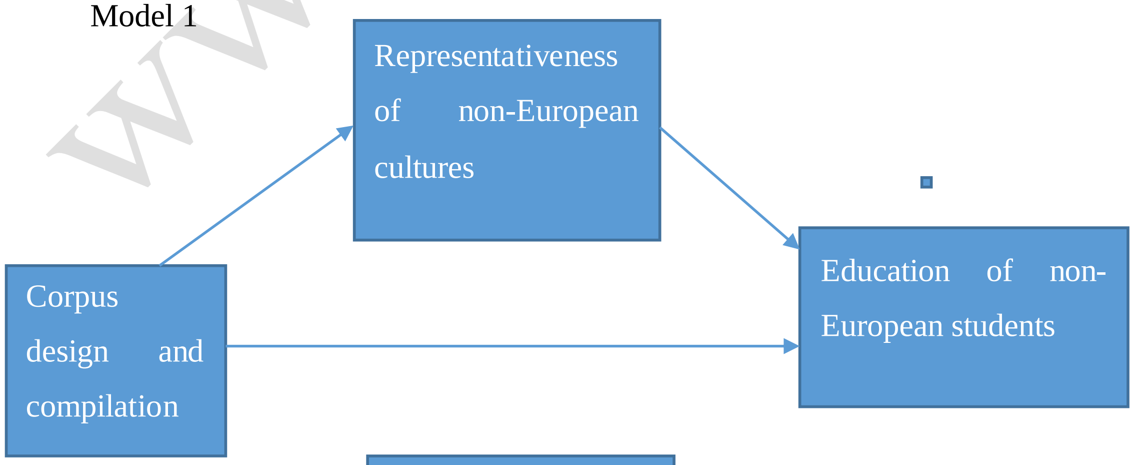
Contemporary education has become culturally, linguistically, and ethnically inclusive to a large extent (Banks, 2019). Facilitation of distance-learning programs and internet-based translations facilities students in continuing their education in overseas institutions with unprecedented ease (Kant, 2021). Due to this transformation, languages are also experiencing a constant evolution due to the interaction of diverse cultures. In this regard, a well-synthesised corpus is aimed to act as a holistic platform to understand the language included in it. Due to this, an effective corpus has to be a representative of all the types of languages and dialects included in it (Reppen, 2019). However, representativeness of a corpus is a complex issue because of different criteria of representativeness. While most modern corpus are representatives of distinct registers such as fictional and nonfictional content, conversations, and speech, considerably fewer are representative of the demographics of the authors (Leech, 2014). Demographic aspects such as nationality, race and ethnicity, cultural background, and native language have been found to influence the studies done through the corpus (Rogers et al., 2021). Corpus based translation studies heavily rely on the inclusiveness of the corpus for local variations in languages and dialects of the same language. This dependence is due to the dynamic nature of languages which also reflect the societal transformations that occur on a continual basis. In the UK, around 605,130 international students were registered in 2020-21, amongst which, around 452,225 belonged to non-European backgrounds (HESA, 2021). The proportion of international students in the UK is witnessing a steady increase of approximately 8.71% annually (HESA, 2021). This increasing figure shows that Corpus linguistics needs to account for an increasingly diverse UK population.

H3: Representativeness of non-European cultures in the corpus significantly mediates the relationship between corpus design and education of non-European students in the UK

H4: Representativeness of non-European cultures significantly mediates the relationship between corpus sampling and education of non-European students in the UK

Conceptual framework

Model 1


Representativeness of non-European cultures
Corpus design and compilation
Education of non-European students

Representativeness of non-European cultures

Education of non-European students

Methodology

Research philosophy is categorised into two types namely; positivism and interpretivism (Saunders et al., 2015). In the current research, researcher has prioritised the philosophy of positivism, as the current research is subjective to factual knowledge, and researcher determined the impact of corpus linguistic impact on education in the UK for non-European students. Another justification for using a philosophy of positivism, it facilitates the researcher in gathering appropriate information while considering the subjective viewpoints in each aspect (Babones, 2016). Moreover, it also assists the researcher in eliminating biases from obtained information by focusing on objective data and incorporating the fundamental variables.

The research approach refers to the process or assumption employed by a researcher for data collection, analysis, and interpretation (Woiceshyn & Daellenbach, 2018). However, there are two main approaches that are widely prioritised by the researcher namely; the deductive approach and the inductive approach. In this particular research, the researcher's aim is to explore the impact of corpus linguistics on education in the UK for non-European students. Therefore, based on the specified topic and nature of the research, a researcher has deployed a deductive approach. The justification for using this approach is as it facilitates the researcher in addressing the research objectives, based on logical reasoning (Zalaghi & Khazaei, 2016). In addition, as per Azungah (2018), the deductive approach is mostly suitable for quantifiable studies, and it assists the researcher in formulation of new theories, based on the existing literature/data.

For research design, the researcher has particularly relied on the quantitative research method, based on the nature of the current research. Along with the philosophy of positivism, the quantitative research method assists the researcher to analyse the impact of corpus linguistics on education in the UK and testing the hypothesis by using statistical tools/techniques (Tobi &

Kampen, 2018). Moreover, it also facilitates the researcher to minimise external errors and increase the validity of outcomes through mitigating psychological and systemic biases.

Referring to the source or steps for data collection, author has collected information using a primary approach, based on the survey questionnaires. Since the current research aim is to analyse the influence of corpus linguistics on education in UK, particularly on Non-European students henceforth, the primary research approach will enable the researcher to collect the most relevant and up-to-date information to carry out this research. For sampling, convenience sampling has been deployed, as per the convenience and accessibility of the researcher. For sample size, 100 non-European students were selected from the UK, as most of the statisticians agree that the minimum sample size for meaningful result is determined to be 100 (Lakens, 2022). The author further added that if population size is found to be less than 100 then researcher need to conider whole population. Therefore, sample size of 100 is considered to be sufficienct to provide a valuable information and desired outcome. Responses were collected on different aspects of the variables to determine the interdependencies, based on the 5-point Likert scale. After collecting data, average has been taken of different question, and SPSS software was used for data analysis, in which correlation analysis, regression analysis, and mediating effect were used to decipher the influence of corpus linguistics on education in the UK for non-European students.

Lastly, ethical consideration has also been valued in this specified research, as it ensures that participants' information would be protected, and not cause any harm. Moreover, it also increases the validity of the outcomes, as the researcher ensure that all information is based on true and authentic sources, and has been appropriately cited.

Findings and analysis

Descriptive statistics

Descriptive statistics in the current case are provided as follows:

Table 1 Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis

	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic
Corpus design and compilation	100	0	3.75	1.5525	0.80348	0.155	-0.098
Corpus sampling	100	0	4	1.52	1.03293	0.468	-0.857
Representativeness of non-European cultures	100	0	3.67	1.3067	0.99118	0.43	-0.631
Valid N (listwise)	100						

As observable from the above table, the mean values of responses under the three variables are around the values of 1.3 and 1.5. Since this value corresponds to the option of ‘agree’ and ‘neutral’ on the Lickert scale used, it can be asserted that average responses for all three variables have been on the agreement side. Meanwhile, the standard deviation value of corpus and compilation, and representativeness variables have been found to be less than 1, indicative of few outliers for these variables. Whereas, it is marginally greater than 1 for Corpus sampling, which indicates that there are a few outlier responses under this variable. The values of Skewness and Kurtosis of all the three variables are also small, indicating that most of the data remains around the normal line of the distribution curve, with marginal tailing.

Correlation analysis

Statistical tests were conducted on selected variables for deciphering the correlation and statistical impact of the corpus design and corpus sampling on education of non-European students in the UK. The results of correlation test are provided in the following table. As observable, Corpus design and compilation was found to be only moderately but positively associated with the Education of non-European students with a coefficient of 0.698. This means that its effects are not highly pronounced on the effectiveness of education of this study group. Meanwhile, Corpus sampling was found to be strongly and positively correlated with a coefficient of 0.989.

Table 2. Correlations

	Corpus design and compilation	Corpus sampling	Representativeness of non-European cultures	Education of non-European students in the UK

Corpus design and compilation	Pearson Correlation	1	.741**	.766**	.698**
	Sig. (2-tailed)		0	0	0
	N	100	100	100	100
Corpus sampling	Pearson Correlation	.741**	1	.978**	.989**
	Sig. (2-tailed)	0		0	0
	N	100	100	100	100
Education of non-European students in the UK	Pearson Correlation	.698**	.989**	.979**	1
	Sig. (2-tailed)	0	0	0	
	N	100	100	100	100

Regression analysis

Regression analysis was conducted to examine the causal relationship between the independent and dependent variables of the study. Since the study also involved mediation effects of Representativeness of non-European cultures in Corpus linguistics.

For Mediation effect of Representativeness of non-European cultures on relation between Corpus design and Education of non-European students in the UK

In the first step, mediation effect of Representativeness was assessed between Corpus design and Education, as shown in the following sections.

Model: 4

Y: EduUK

X: CorpDes

M: Rep

OUTCOME VARIABLE: Rep

When regression was conducted by taking Representativeness as the outcome variable, R and R-square values show high significance. This indicates that the model itself has a high explanatory and predictive power to explain the relationship model.

Table 3. Summary of model

R	R-sq	MSE	F	df1	df2	p
0.9119	0.8315	0.1672	483.563	1.000	98.000	0.000

However, as shown in the following table from the values of LLCI and ULCI, the model has insignificant explanatory capability to describe the behaviour of Corpus design variable.

Table 4. Model

	coeff	se	t	p	LLCI	ULCI
constant	-0.1625	0.0783	-2.0741	0.0407	-0.3179	-0.007
CorpDes	0.8395	0.0382	21.9901	0	0.7637	0.9153

OUTCOME VARIABLE: EduUK

By taking the Education of non-European students, the following model summary shows high R and R-square with around 99% accuracy for behaviour of the variable.

Table 5. Summary of model

R	R-sq	MSE	F	df1	df2	p
0.9949	0.9899	0.0113	4742.851	2.000	97.000	0.000

Similar to the case above, although the model itself has high predictive capacity, it shows insignificant effects of both Corpus design and Representativeness of non-European cultures.

Table 6. Model

	coeff	se	t	p	LLCI	ULCI
constant	-0.0632	0.0208	3.0361	0.0031	-0.1044	-0.0219
CorpDes	0.4163	0.0242	17.2212	0	0.3683	0.4643
Rep	0.6209	0.0263	23.6473	0	0.5688	0.673

Total Effects Model

From the perspective of total effects, the following model depicts around 93%-96% accuracy of explaining the behaviour of the involved variables.

Table 7. Summary

R	R-sq	MSE	F	df1	df2	p
0.9652	0.9315	0.0757	1333.134	1.000	98.000	0.000

However, as the tables below exhibit, the total, direct, as well as indirect effects of the Corpus design on Education of non-European students was found to be insignificant. These results indicate that education prospects of these students are not significantly influenced by the design and compilation of the corpus.

Table 8. Total effect of independent on dependent variable

Effect	se	t	p	LLCI	ULCI	c_cs
0.9375	0.0257	36.5121	0	0.8866	0.9885	0.9652

Table 9. Direct impact of independent on dependent variable

Effect	se	t	p	LLCI	ULCI	c'_cs
0.4163	0.0242	17.2212	0	0.3683	0.4643	0.4286

Table 10. Indirect impact of independent on dependent variable

	Effect	BootSE	BootLLCI	BootULCI
Rep	0.5213	0.0273	0.4683	0.5773

Table 11. Completely standardized indirect impact of independent on dependent variable

	Effect	BootSE	BootLLCI	BootULCI
Rep	0.5366	0.0228	0.4916	0.5815

For Mediation effect of Representativeness of non-European cultures on relation between Corpus sampling and Education of non-European students in the UK

In proceeding step, the researcher evaluated the mediation effect of Representativeness on the relation between Corpus sampling and Education of non-European students.

Model: 4

Y: EduUK

X : CorpSam

M : Rep

OUTCOME VARIABLE: Rep

By taking Representativeness as the outcome variable, following model summary in table depicts a high ability of the proposed to explain the behaviour of the variables included with statistical accuracy ranging between 95% to 97%.

Table 12. Summary of model

R	R-sq	MSE	F	df1	df2	p
0.9784	0.9572	0.0425	2190.719	1.000	98.000	0.000

This was also supported by the model summary of Corpus sampling alone, showing high significant p-value for the variable.

Table 13. Summary

	coeff	se	t	p	LLCI	ULCI
	-				-	-
constant	0.1203	0.0368	-3.2698	0.0015	0.1934	0.0473
CorpSam	0.9388	0.0201	46.8051	0.000	0.899	0.9786

OUTCOME VARIABLE: EduUK

By taking the variable of Education of non-European students in the UK, the model summary also found high accuracy in the range of 98%-99% to explain the variables' behaviour.

Table 14. Summary of the model

R	R-sq	MSE	F	df1	df2	p
0.9906	0.9813	0.0209	2545.509	2.000	97.000	0.000

The following table shows high significance with respect to the variables, Corpus sampling and Representativeness individually.

Table 15. Model

	coeff	se	t	p	LLCI	ULCI
constant	-0.0108	0.0272	-0.399	0.6907	-0.0647	0.0431
CorpSam	0.7319	0.0679	10.7739	0	-0.5971	0.8667
Rep	0.287	0.0708	4.0545	0.0001	-0.1465	0.4275

TOTAL EFFECT MODEL

OUTCOME VARIABLE: EduUK

With respect to the direct and indirect effects, the researcher found high accuracy of the model summary, as elaborated in the following tables with respect to the Corpus Sampling.

Table 16. Summary of the model

R	R-sq	MSE	F	df1	df2	p
0.989	0.9781	0.0242	4383.92	1.000	98.000	0.000

Table 17. Model

	coeff	se	t	p	LLCI	ULCI
constant	-0.0454	0.0277	-1.6354	0.1052	-0.1004	0.0097
CorpSam	1.0013	0.0151	66.2112	0	-0.9713	1.0314

Meanwhile, the following tables of total, direct, and indirect effects of Corpus sampling on Education of non-European students demonstrate a high significance in both the direct and indirect effects cases through the values of LLCI and ULCI. This means that Corpus sampling significantly impacts the Education prospects of non-European students in the UK directly as well as through the mediation effects of Representativeness of the corpus.

Table 18. Total impact of independent on dependent variable

Effect	se	t	p	LLCI	ULCI	c_cs
1.0013	0.0151	66.2112	0	- 0.9713	1.0314	0.989

Table 19. Direct impact of independent on dependent variable

Effect	se	t	p	LLCI	ULCI	c_cs
0.7319	0.0679	10.7739	0	- 0.5971	0.8667	0.7229

Table 20. Indirect impact of independent on dependent variable

	Effect	BootSE	BootLLCI	BootULCI
Rep	0.2695	0.0675	-0.1536	0.4175

Table 21. Completely standardized indirect impact of independent on dependent variable

	Effect	BootSE	BootLLCI	BootULCI
Rep	0.2661	0.0657	-0.153	0.4094

Discussion and Hypothesis Summary

This research has been carried out to determine the impact of corpus linguistic on education of non-European students in UK. Referring to the impact of corpus design and compilation, the literature had pointed out a substantial role of course design and compilation in providing a quick access to English vocabulary, and facilitating student and teacher through collaborative environment (Kogan et al., 2020; Li et al., 2022). Similarly, it has also been substantiated in the current research, as findings in the current study revealed that higher association between corpus design and compilation and education of non-European students in UK. Further, findings from regression analysis has also shown that corpus design has a significant and positive influence on the education of non-European students in UK. Thus, hypothesis 1 is found to be correct, and it implies that corpus design and compilation has a significant impact on education of non-European students in the UK.

On contrary, Corpus sampling has also been used to analyse its influence on the education of non-European students in UK. Findings in the previous studies revealed that corpus sampling plays an essential role in enhancing the breadth of knowledge about a particular foreign language for non-native speakers (Zahra & Abbas, 2018). Additionally, it also facilitate the learners to in gaining both, the experience of contributing to a foreign language and the confidence for future research in that foreign language (Kogan et al., 2020). Similarly, findings from correlation analysis has also been revealed that there is a positive and strong association between corpus sampling and education of non-European students in the UK. Further, findings from regression analysis has also shows that corpus sampling has a significant and positive influence on the education of non-European students in the UK. Thus, findings in the current study are in line with previous studies, and hypothesis 2 has also been accepted, which suggested that corpus sampling has a significant impact on education of non-European students in the UK.

Further, representativeness of non-European cultures has also been used to analyse its mediating effect on education of non-European students in the UK. Henceforth, hypothesis 3 was found to be incorrect, as representativeness of non-European culture was found to be playing insignificant role between corpus design and education of non-European students in the UK. Meanwhile, hypothesis 4 was found to be correct, as representativeness of non-European culture was found to be playing significant mediating role between corpus sampling and education of non-European students in the UK.

Table 22. Hypotheses summary

S.No.	Developed and tested hypotheses	Status
1	Corpus design and compilation has a significant impact on education of non-European students in the UK	Accepted
2	Corpus sampling has a significant impact on the education of non-European students in the UK	Accepted
3	Representativeness of non-European cultures in the corpus significantly mediates the relationship between corpus design and education of non-European students in the UK	Rejected
4	Representativeness of non-European cultures significantly mediates the relationship between corpus sampling and education of non-European students in the UK	Accepted

Conclusion

It was discovered that corpus design and compilation were only marginally yet positively related to non-European students' education. This indicates that its impacts on the success of this student group's schooling are not highly evident. Corpus sampling, on the other hand, was discovered to have a 0.989 coefficient of strong and positive correlation. However, it was determined that the overall, direct, and indirect effects of the Corpus design on the education of non-European students were negligible. These findings suggest that the design and compilation of the corpus did not significantly affect the educational prospects of the non-native students of linguistics. Meanwhile, the researcher found a significant impact in both the direct and indirect effects situations for the impacts of corpus sampling on the education of non-European students. This suggests that both directly and indirectly through the consequences of the corpus' representativeness, corpus sampling has a considerable impact on the educational chances of non-European students in the UK.

Future implications

Since the current study examined the influence of Corpus design and sampling on the effectiveness of education of non-European linguistic students in the UK, it has far-reaching implications for the academia. The findings will facilitate future researchers in focusing the specific aspects of corpus sampling such as variations in words and their lexical aspects.