Corpus linguistics in building professional communicative competencies of graduates of higher educational institutions

Abstract: The purpose of the article is to study the role of corpus linguistics methods in the formation of students’ professional communicative competence. The study used the method of expert evaluation, a method of questionnaire with open questions. There was a change in the minimum score in the experimental group ranging from 50 to 67. Such results are significant and demonstrate the effectiveness of the method under consideration. After the experimental training, the value of the criterion was $T_{emp} = 4.41$, which is higher than the critical value. The practice of creating a corpus at the stage of course planning and selection of educational content has proven its capability and, most importantly, its versatility. The use of both the idea of compiling a subject area corpus and the statistical evaluation of vocabulary is highly effective. A scientifically based selection of professionally relevant foreign language vocabulary items is effective in teaching a foreign language for any profession. Further research should be aimed at expanding the scope of corpus linguistics in this area.

Keywords: English language competence, corpus linguistics, innovative technologies, higher education, Cambridge learner corpus.
significant didactic potential in the development of methods of working with vocabulary, but it is not sufficiently used to develop communication skills (Ishchenko et al., 2022).

Any form of communication is unique and requires communication among participants to have a high level of communicative competence. This is necessary to form a goal and find an optimal way out of the situation, to be tolerant of each other’s social and cultural differences, etc. The methods and tools of corpus linguistics are increasingly used as an innovative method of building communicative competencies (Dilai & Dilai, 2021).

Corpus linguistics is one of the most rapidly developing methods in modern theoretical and applied linguistics. In linguistics, a corpus is a collection of texts carefully selected and processed according to certain rules, which is used for statistical analysis and hypothesis testing, confirmation of certain linguistic rules in a given language (Paziljanova, 2022).

The analysis and processing of different types of corpora is the subject of most studies in computational linguistics (e.g., keyword extraction), language recognition, and machine translation. The corpora are currently often used to create hidden tagged models for part-of-speech tagging, and other tasks (Römer, 2022).

In view of the foregoing, it is important to take into account the semantic difference between the concepts of “communicability” and “sociability” when forming a conscious attitude of students to the competencies they should have (Solan, 2020). Accordingly, these concepts should be differentiated with respect to competencies in educational standards given the relevance of this issue (McEnery et al., 2019).

Having analyzed a number of existing models of communicative competence (Savignon, 1997; Hutmache, 1997; Ek & Trim, 1998), it is possible to conclude that communicative competence is characterized by three components — value, content, and structure. For the most part, researchers have identified three to six components of communicative competence (Figure 1).

**FIGURE 1** - Component of communicative competence models before and after the training

![Diagram of communicative competence models](source: created by the authors)
Thus, the models of the above-mentioned authors have similar components:
- the ability to use the whole set of phonetic, lexical, grammatical rules and linguistic means for the correct construction of statements (Stefanowitsch, 2020);
- the ability to formulate thoughts using language means for the purpose of communication, the ability to choose the appropriate form of speech depending on the purpose and situation of communication (Xodabande & Nazari, 2022b);
- the ability to interact adequately in situations of everyday life and professional activity, the willingness to maintain and develop contacts, taking into account the communicative norms of communication in different cultures (Asención-Delaney et al., 2022).

In practice, there is a contradiction between the requirements of employers and the level of foreign language communicative competence of graduates. To overcome this, it is necessary to seek an appropriate level of graduates’ communicative competence that makes them ready for their new professional requirements (Xodabande & Nazari, 2022a). This should be attained in accordance with the requirements set out in professional profiles that determine the methods chosen for their development. It not only combines general cultural competencies but also acts as an important element in the development of professional competence (Alimjanova, 2022).

Unexplored issues

With the development of society, the need for training specialists is constantly changing, which inevitably comes into conflict with the existing education system. In this regard, it is necessary to constantly adjust the existing teaching methods, improve the processes of selecting the content of educational programs, and develop new teaching materials. In particular, the foreign language programs used to train future specialists have significant shortcomings. The development and realization of the linguistic and didactic potential of language corpora will help to correct the discrepancies that constantly arise between the existing system of foreign language training and the validity of its results.

The aim

The aim of the study is to confirm the effectiveness of corpus linguistics in the development of professional foreign language communicative competence of EFL students in HEIs.

Objectives

1. To study the influence of the use of corpora on the process of building professional communication competencies during the study of EFL in HEIs.
2. To create educational corpus-based material for the English language course.
3. To identify statistically significant differences between the academic performance of the control and experimental groups.

Literature review

A corpus is currently used to solve tasks that were previously impossible. Besides, not only the efficiency and speed of language data processing but also their reliability have increased over time. In the “pre-corpus era”, the accepted alternative was always considered intuition, which was not always supported by sufficient evidence. As Holoshchuk (2022) stated, the teaching of a foreign language is in a constant search for ways to transfer experience in the most effective way. Corpus tools, as a means of managing a large amount of data in various ways, create a chaotic variety of languages (Zhydenko, 2021). Combined efforts of corpus linguistics and linguistic didactics will result in finding an adequate and proportional reflection of the limitless element of language in the available limited-scope of educational material.

The main attention of the researchers was focused on the knowledge of the phenomena of human language, the existence of the word in the context, rather than the abstract logical reasoning.
Farr et al. (2022) selected not only the study of the regularities of the language as a priority but also their verification in practice. The authors studied the combination of words based on the unity of collocation and colligation, as the correct construction of the language is impossible if only grammatical or collational reliability is observed.

Summarizing findings from corpus studies, Callies (2019) argues that sometimes clear dichotomies facilitate understanding but impoverish reasoning. The author emphasizes the need to consider some phenomena in terms of a scale, or a spectrum that allows for intermediate options. Some concepts may be even more complex and require a multidimensional model.

Çyfeku (2022) focuses on the transition from written to spoken language, from speaking to listening, and from product to process. The path from knowledge to skills, from structure to vocabulary should be based on the integration of corpus linguistics methods and foreign language learning technologies. According to the author, this process is capable of leading to the creation of effective means of developing the communicative competencies of students of HEIs.

As Zhurabekova (2022) states, diversity of speech is achieved by assimilating a large number of fixed and semi-fixed, perceived and reproduced ready-made units. One cannot disagree with the fact that a generative, creative potential is necessary for generation, which enables the formulation of something that has never been expressed before. According to the author, this is emphasized in traditional approaches with a bias toward grammar.

Nazarova (2022) believes that the lexical approach implements a broader view of vocabulary, which is related to the reality of speech, the study of lexical units. The researchers state that this does not imply opposition between vocabulary and syntagmatics, but their correlation. The boundaries and spheres of influence of vocabulary are expanding, which corresponds to the current state of linguistic sciences.

A corpus format can also help to avoid the right/wrong dichotomy of vocabulary in English classes. According to Liu et al. (2022), it enables replacing this polarity with a spectrum of probable usage of an utterance, based not on feelings, but on quantitative regularities. The authors consider this approach more adequate in the context of learning a foreign language in HEIs.

It should be noted that Li (2022) considered the transition from a single word to a word combination as the actions of two processes in their unity. Considering the word, only grammatically is regarded insufficient.

According to Kizi Tolmasova (2022), word combinations and phrases are made of units of meaning. If the word is independent and has an independent meaning, it is most likely on the border of linguistic interest. Such words include the lists of representatives of flora and fauna.

Based on numerous studies on this topic, we can conclude that corpus linguistics is playing an increasingly important role in EFL. Despite the lack of sufficient research on the effectiveness of corpus linguistics methods in the process of developing foreign language competencies, its role cannot be ignored. This study aims to fill the existing gaps and consider the practical aspect of using corpora of educational texts.

**Methods**

**Design**

The article summarizes the results of the research, which was conducted in three stages:

1. The first stage (2021): At this stage, the goal and objectives of the study were organized. The concept for the study of the use of corpus linguistics in building professional communication competencies in HEIs was also developed. At this stage, the Cambridge Learner Corpus (CLC) was chosen as the learning corpus.

2. The second stage (2022) involved the definition of technologies for the development of students’ professional communication competencies using corpus linguistics. A sample was for-
med, and a pedagogical experiment - the use of corpora for the development of foreign language professional communication competencies — was conducted.

3. The third stage (2023) had to do with the research results. A statistical analysis of the results was carried out and conclusions were drawn.

**Participants**

A draw method was used for sampling from the general population, which was implemented in several stages. In the first stage, all elements of the general population were marked. In the second stage, the required number of cards was selected from the deck at random. These cards were set aside and did not participate in the further selection. The nested sampling was carried out in this way. The nested sample involved the selection of several courses from the general population, within which the survey was conducted in a continuous manner. The study of the effectiveness of corpus linguistics methods in building professional communication competencies was conducted at the Faculty of Pedagogical Education of Borys Grinchenko Kyiv University. The survey involved a total of 186 students in the 2nd-4th year of the Faculty of Pedagogical Education majoring in Psychology, Choreography, and Visual and Musical Arts. Such a sample covers the required number of respondents to ensure the reliability of the results obtained. The respondents were divided into two equal groups — control and experimental. Corpus linguistics tools were used in the experimental group during English language learning.

**Instruments**

The participants took part in the study through an online survey, which was carried out by using Google Forms. Cambridge Learner Corpus (CLC) contains exam scenarios created by more than 180,000 students around the world and is constantly updated. The corpus manager Wordsmith 5.0 was used to work with the corpus.

**Data collection**

1. The principles of selecting textual material for a methodologically oriented corpus should, first of all, be consistent with didactic and methodological principles, since the main purpose of compiling such a corpus is to identify and teach professional vocabulary. Some of the texts are used as language material for learning. Within the communicative approach, purely communicative exercises are described:

2. The method of expert evaluation encompassed a group of 20 experts from among the professors of the Faculty of Pedagogical Education to identify the level of professional communicative competence of the university students;

3. The method of open questioning in the discipline was used to monitor students’ academic performance. The tasks for students of both groups were the same and corresponded to the curriculum. Cronbach’s alpha coefficient is 0.69.

Statistical processing of samples of the control and experimental groups to assess the reliability of the obtained data was carried out according to several criteria only at the significance level of 0.05, which corresponds to the reliability level of 95%. The following hypotheses were advanced:

– the hypothesis of no differences (null hypothesis);
– the hypothesis about the significant differences (alternative hypothesis).

**Analysis of data**

1. The chi-squared criterion was calculated using the formula:

\[ \chi^2 = \frac{(f_1 - f_2)^2}{f_1 + f_2} \]  \hspace{1cm} (1)

where \( f_1 \) and \( f_2 \) – frequencies of compared samples.

2. The Cronbach’s alpha reliability coefficient indicates the internal consistency of the test assignments. The Cronbach’s alpha coefficient is calculated using the formula:

\[ \frac{N}{N-1} \left( 1 - \frac{1}{\sigma^2} \left( \frac{1}{N} \sum_{i=1}^{N} \sigma^2_i - \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2 \right) \right) \]  \hspace{1cm} (2)
where $\sigma_2^2$ – total test score variance; 
$\sigma_i^2$ – variance of i element.

3. The Cramer-Welch criterion:

$$T_{emp} = \frac{\sqrt{MN}(|x-y|)}{\sqrt{DX + D_Y}}$$  \hspace{1cm} (3)

where X – the experimental sample (corresponds to the experimental group), Y — the control sample (corresponds to the control group), N — the volume of the experimental sample, M — the volume of the control sample, $D_X$, $D_Y$ — sample variances.

4. The calculation of the Mann-Whitney U test is calculated using the formula:

$$U = (n_1n_2) + \frac{(n_1n_1+1)}{2} - T_x$$  \hspace{1cm} (4)

where $n_x$ – the number of respondents in the experimental group; $n_2$ – the number of respondents in the control group; $T_x$ – the larger of the two rank sums; $n_x$ – the number of respondents in the group with a higher sum of ranks.

**Ethical criteria**

All respondents were asked in advance to independently and honestly answer the questions posed during the survey. The students were introduced to the pedagogical conditions — the means of corpus linguistics in building professional communicative competencies. The respondents themselves, as well as the questionnaires used, included ethical demands of integrity, competence, respect for the individual, and anonymity. The objectivity and impartiality of the research results cannot be doubted. The methodological imperfection of research tools and means are the main limitations of the study. There are no direct methods for finding the connection between the use of corpus linguistics and the level of communicative competencies.

**Results**

Didactic understanding of the potential of corpus technologies provides ample opportunities both for designing the content of learning and for creating learning models that better meet the current objectives of HEIs. The content of teaching practical professional-oriented foreign language skills is actually a set of things that students should master in order to achieve a level of foreign language skills that would enable them to solve the problems of everyday professional communication.

The pedagogical experiment was conducted in order to empirically confirm the proposed research hypothesis. This conclusion is made on the basis of statistical processing of the data obtained as a result of the experiment. The value of the empirical criterion is calculated for this purpose using the data obtained before and after the experiment for the control and experimental groups. Table 1 presents the numerical dynamics of academic performance.

**TABLE 1 - RESULTS OF ACADEMIC PERFORMANCE MEASUREMENTS IN THE CONTROL AND EXPERIMENTAL GROUPS AT THE BEGINNING AND AT THE END OF THE EXPERIMENT**

<table>
<thead>
<tr>
<th>Item No.</th>
<th>CG at the beginning</th>
<th>CG at the end</th>
<th>EG at the beginning</th>
<th>EG at the end</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67</td>
<td>70</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>2</td>
<td>61</td>
<td>72</td>
<td>78</td>
<td>83</td>
</tr>
<tr>
<td>3</td>
<td>59</td>
<td>68</td>
<td>60</td>
<td>67</td>
</tr>
<tr>
<td>4</td>
<td>71</td>
<td>64</td>
<td>68</td>
<td>76</td>
</tr>
<tr>
<td>5</td>
<td>79</td>
<td>68</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>71</td>
<td>79</td>
<td>63</td>
<td>71</td>
</tr>
<tr>
<td>7</td>
<td>58</td>
<td>66</td>
<td>65</td>
<td>71</td>
</tr>
<tr>
<td>8</td>
<td>58</td>
<td>72</td>
<td>62</td>
<td>73</td>
</tr>
<tr>
<td>9</td>
<td>62</td>
<td>63</td>
<td>61</td>
<td>70</td>
</tr>
</tbody>
</table>
Next, it is necessary to compare the value of the obtained criterion with the critical one. If the empirical value of the criterion is below the critical one, it can be claimed that the characteristics of the two studied groups coincide. Otherwise, the indicators of the groups are different. Table 2 provides calculations of the reliability of experimentally obtained data.

**TABLE 2 - STATISTICAL CALCULATIONS FOR SAMPLES OF CONTROL AND EXPERIMENTAL GROUPS**

<table>
<thead>
<tr>
<th></th>
<th>CG (1)</th>
<th>CG (2)</th>
<th>EG (1)</th>
<th>EG (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arithmetic mean</td>
<td>63.40</td>
<td>67.57</td>
<td>62.66</td>
<td>76.88</td>
</tr>
<tr>
<td>Standard error of the mean</td>
<td>1.48</td>
<td>1.52</td>
<td>1.70</td>
<td>1.44</td>
</tr>
<tr>
<td>Median</td>
<td>62</td>
<td>68</td>
<td>62</td>
<td>78</td>
</tr>
<tr>
<td>Mode</td>
<td>59</td>
<td>70</td>
<td>61</td>
<td>78</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>5.15</td>
<td>5.20</td>
<td>6.61</td>
<td>5.72</td>
</tr>
<tr>
<td>Variance</td>
<td>26.62</td>
<td>28.22</td>
<td>42.80</td>
<td>32.81</td>
</tr>
<tr>
<td>Sample size</td>
<td>14</td>
<td>19</td>
<td>28</td>
<td>21</td>
</tr>
</tbody>
</table>

**SOURCE:** CREATED BY THE AUTHORS

However, some other indicators of descriptive statistics of samples, such as the minimum score and the interval of values, should be noted. The minimum score in the control group before and after the experiment almost did not change. It was 58 before the experiment, and it became 62 after the experiment. In the experimental group, the minimum score increased significantly: from 50 to 67.

The empirical value of the criterion is based on the results of the first test before the experiment: \( T_{emp1}=0.24 \), which is less than critical. This result means that the null hypothesis about the absence of differences between the experimental and control groups is accepted. After experimental training the value of the criterion is \( T_{emp2}=4.41 \), which is more than critical. Therefore, the alternative hypothesis that the reliability of a positive increase in the measured indicator of the control and experimental groups after the experiment is 95% shall be accepted. Table 3 presents the results of the comparison by other criteria.

**TABLE 3 - COMPARISON OF TEST RESULTS IN CONTROL AND EXPERIMENTAL GROUPS USING CRA-MER-WELCH AND WILCOXON-MANN-WHITNEY TESTS**

<table>
<thead>
<tr>
<th>Cramer-Welch test</th>
<th>Y before</th>
<th>X before</th>
<th>0.23</th>
<th>The characteristics of the compared samples coincide only at the significance level of 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y after</td>
<td>X after</td>
<td>4.40</td>
<td>The reliability of the differences in the indicators of the compared samples is 0.95</td>
</tr>
</tbody>
</table>

**SOURCE:** CREATED BY THE AUTHORS
Wilcoxon-Mann-Whitney test 1.96

Y before X before 0.33 The characteristics of the compared samples coincide only at the significance level of 0.05

Y after X after 3.31 The reliability of the differences in the indicators of the compared samples is 0.95

SOURCE: CREATED BY THE AUTHORS

So, it can be argued that the differences in the level of professional communicative competencies obtained as a result of the experiment in the experimental and control groups cannot be explained by random factors only. It follows that the experiment confirms the higher efficiency of using corpus selection of professional vocabulary and corpus-based learning in comparison with traditional methods. A survey after the end of the English language course revealed that, despite the difficulty of working with non-adapted texts and lexical data, 99% of the subjects noticed a positive effect from such activities and expressed a desire to continue using the “live” language materials of the corpus, as well as films, analytical programmes and articles and Internet materials.

A desire was expressed to increase the number of current linguo-methodical materials developed by the university (foreign language department) and related to the future profession. All participants of the experiment also expressed a desire to speak, listen, and read more in English classes.

The students indicated the following effective methods of practicing vocabulary: “studying examples of the corpus and dictionary articles”, “exercises to practice the word in all connections, grammatical forms”, “translating sentences from Ukrainian to English and vice versa”. Subjectively, some advantage of the tested experimental groups was also a great tolerance for unfamiliar words in the text.

The increased responsibility of students for the result of their studies, the willingness to actively participate in the process of acquiring knowledge, skills, and abilities as opposed to the passive expectation of ready-made knowledge from the teacher should also be noted. The general comparison of the results of the experiment, i.e., skills and abilities, including educational ones, obtained as a result of experimental training, speaks about the rationality of the proposed methods of teaching vocabulary to students of the specified majors.

As the samples of students in the study are random and independent, the measured property had a continuous distribution and was examined by experts in four categories (creative, productive-creative, constructive, reproductive), so the two-sided criterion $\chi^2$ can be used. Using the statistical table for the critical values of statistics having the $\chi^2$ distribution, it was found that for the significance level $\alpha = 0.05$ and the number of degrees of freedom $n-4-1=3$, the value of $T_{\text{crit}}$ was 7.81. It follows from the results of the experiment that $T>T_{\text{crit}} (13.92>7.81)$. This gives grounds for accepting the null hypothesis. Therefore, the values of the measured $O_1$ and $O_2$ differ significantly from each other for all $i=1, 2, 3, 4$. So, at the end of the experiment, the percentage of students with high and medium levels of indicators is significantly higher in the experimental groups, and the percentage of students with low levels of indicators is significantly lower than in the control groups. This is why it can be argued that the consequence of the implementation of pedagogical conditions is an increased level of professional communicative competencies.

Discussion

The results of the data obtained indicate the high efficiency of using corpus linguistics methods in the development of foreign language communicative competence of university students. The data obtained from the respondents of the experimental and control groups indicate a significant impact of the corpus of educational texts on the development of communicative competence. Thus, it should be noted that there
is a statistically significant difference in the results of the students of the experimental group. This indicates that the use of corpus linguistics tools is more effective than traditional teaching methods (results of the control group).

As Horokhova (2022) noted in her study, corpora and frequency dictionaries can be useful in foreign language learning. The authors propose to classify corpora according to various features: the purpose of creating a corpus, the type of linguistic data, ‘literariness’, genre, dynamism, mark-up type, volume of texts, etc. For example, according to the parallelism criterion, corpora can be divided into monolingual, bilingual and multilingual. According to the authors, such corpora can be used in building students’ communicative competencies. Soliyev (2022), emphasizes that communicative competencies can be built in the process of communicative practice only.

In order to achieve professional success and career growth for future specialists, it is necessary for a graduate to have a level of communicative competence that meets the employers’ requirements. A number of authors, in particular Jurabekova (2022) and Shestopalova (2021) emphasize that these competencies currently play an important role for employers. Graduates who do not know how to build phrases and sentences correctly and logically, correctly convey their opinion to the interlocutor, are not ready for professional communication. As the results of this study show, it is the tools of integrated linguistics that help solve this problem. According to the researchers, students do not know the rules of behavior in a business environment and do not have the skills for public presentation of the results of their work. Musratova (2022) notes that despite the sociability characteristic of graduates, employers are convinced that graduates do not know how to conduct professional communication. They also do not have the ability to reach a compromise, do not know how to be flexible in difficult situations, are not familiar with the technique of social interaction and cannot get along with people of different cultures.

Xodabande & Nazari (2022b) emphasized the role of vocabulary in their studies, noting that words are names of things, phenomena, events, and names of everything that is and can be in reality. If the name is incomprehensible to the listener, it is empty — no such object exists. Instead, Solan (2020) states that the issue of building communicative competencies is in the core of general professional competence. Römer (2022) described the role of corpus linguistics in building professional communicative competence. The author emphasizes the great value of text corpora for developing various professional communication strategies.

As the results of this study show, it is the tools of integrated linguistics that help solve this problem. This makes it possible to assert the effectiveness of corpus linguistics in the development of foreign language communicative competencies in higher education students.

The theoretical significance of the research is that the work, along with the theoretical justification of the need to introduce corpus resources into the educational process, presents their didactic understanding. In the experiment, linguistic didactic methods of working with the corpus were applied, the didactic concepts of the unit of study, corpus competence, research task were defined, and a model was created. The practical value of the work lies in the fact that the considered corpus can be used to select keywords, develop a professional lexical corpus based on them. The materials of the corpus are a source of examples of the professional content of the development of tests and exercises. The experience of compiling and using the CLC will be of interest to those who develop similar problems related to teaching a professionally oriented foreign language in other settings. The main limitations of the study were the lack of a methodology for studying the impact of using corpus methods in the process of training future specialists for professional communication.

**Conclusions**

The relevance of the study is determined by the need to find effective means of building
Communicative competencies in students of HEIs. It should be noted that the introduction of corpus methods into the educational process has a positive effect on the assimilation of professional vocabulary, which is reflected in the results of the experiment. An important result is an increase in the professional competence of the participants of the experimental group. The research assignments aimed at teaching and studying the features of professional language communication based on the corpus tools were introduced in the course of testing the research hypothesis. Based on the results of their application during the lessons, it can be concluded that the corpus and corpus methods enable both analysing and obtaining the most representative linguistic material. This made it possible to develop effective tasks and exercises for consolidating lexical units of a professional direction, assimilating the peculiarities of their use in speech. The results obtained during the research may be of interest to both employees of educational institutions and students who strive to build communicative competencies. The research will be useful in the creation of foreign language training programmes focused on building students’ communicative competencies in HEIs. The research may also be of interest to corpus linguistics specialists. The obtained results can also be used in lecture courses on the theory and methods of teaching foreign languages. Further research should be aimed at studying the use of corpus linguistics tools for educational purposes: for building other foreign language competencies.

References


McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus linguistics, learner corpora, and SLA:


