

УДК 811.111'01/01

DOI: <https://doi.org/10.32589/2311-0821.1.2023.286180>

O. Iu. Andrushenko

Kyiv National Linguistic University, Ukraine

e-mail: and.olenka@gmail.com

ORCID ID: <https://orcid.org/0000-0002-7699-9733>

LANCSBOX SOFTWARE OPTIONS FOR THE PROSPECTIVE INVESTIGATION OF THE MULTILINGUAL CORPUS FOR EUROPEAN STUDIES

Abstract

The paper presents a comparative analysis of the lexeme *European* in two language variations (British and American English) based on the built-in corpora represented by newspapers, fiction, etc. that are licensed by LancsBox software (AmE06 and BE06 respectively). The investigation describes the algorithms of implementing linguistic research as part of the project taught during the course “Multilingual Corpus and its Resources for European Studies (KNLU)” (Erasmus+ Program). The LancsBox user-friendly software, that works with major operating systems, has proved to be a powerful manager for compiling and using the existing corpora. It enables to visualize the textual data based on the following software package tools: KWIC, GraphColl, Words, Ngrams, Wizard, etc. essential for the study of a specific linguistic unit. The statistical analysis of both corpora under analysis has revealed that the word *European* belongs to the lexemes that are seldom employed in the language. The comparison of the two variations has shown that the word occurs in similar top-ten frequent collocates, however, the GraphColl tool visualization has indicated the major differences between two corpora. Thus, in British English Corpus N+N structures are more commonly employed and are more vibrant than in American English Corpus. The t-test has proved a statistically significant difference between the corpora with regard to the linguistic variable *European*. These data may testify to cultural differences between the users of two language variations taking into account that both corpora represent the same time frame.

Keywords: *European*, LancsBox, corpus studies, corpus tools, automated analysis.

Анотація

У статті представлено компаративний аналіз лексеми *European* у двох мовних варіантах англійської мови (британський та американський) на основі вбудованих ліцензованих корпусів програмного забезпечення LancsBox (AmE06 та BE06 відповідно), що репрезентовані газетними статтями, художньою літературою тощо. Описано алгоритм виконання лінгвістичного дослідження, що є частиною проекту “Мультилінгвальний корпус та його ресурси для дослідження Європеїстики” (КНЛІУ) (програма Erasmus+). LancsBox – зручне програмне забезпечення, що працює з основними операційними системами та є ефективним менеджером для укладання й використання вже наявних корпусів. Це дає змогу візуалізувати текстуальні дані на основі наступного пакету програмного забезпечення: KWIC, GraphColl, Words, Ngrams, Wizard. Вони є основними для вивчення окремої лінгвістичної одиниці. Статистичний аналіз обох окреслених корпусів довів, що слово *European* належить до лексичних одиниць із низькою частотою використання в мові. Порівняння двох мовних варіантів показало, що слово використовується в майже однакових найпоширеніших 10 колокаціях, проте при імплементації інструмента візуалізації GraphColl зауважена основна відмінність між уживанням одиниць в корпусах. Так, у корпусі британської англійської мови найчастіше трапляються структури N+N, що більш динамічні порівняно з відповідними структурами в корпусі американської англійської мови. Окрім цього, Т-тест статистично показав значну різницю між корпусами у функціонуванні лінгвістичної змінної *European*. Отримані дані можуть свідчити про культурну відмінність носіїв у двох мовних варіантах, зважаючи на те, що обидва корпуси представляють тексти, укладені в межах однакових часових рамок.

Ключові слова: *European*, LancsBox, корпусні дослідження, корпусний інструментарій, автоматичний аналіз.

Introduction. Implementing the project “Multilingual Corpus and its Resources for European Studies Research (KNLU)” the article presents the LancsBox tool (Brezina et al, 2020) and its options for the automated analysis of built-in and self-compiled corpora which enable the investigation of a specific search term with reference to the language selected. The course aimed at PhD students of Kyiv National Linguistic University is carried out within Jean Monnet Activities (Erasmus+ Program) and has a goal to provide practical instruments for the young researchers to conduct their linguistic research. The **aim** of the current paper is to study the lexeme *European* in two balanced corpora (American English and British English) built-in LancsBox software that contain 500 texts each (Baker, 2009; Potts & Baker, 2012) and present software opportunities for the future automated analysis of the Multilingual corpus for European studies.

Literature review. Recent developments in corpus linguistics and the relevant technological progress have enabled to elaborate specific software for analysing the language. Such software appears to be more intuitively friendly for scholars who are not experts in computer science (O’Keeffe&McCarthy, 2021). The access to the corpora via online interfaces has “empowered a broader number of linguists to explore the data from a greater range of languages, which wasn’t the case in the last decade, providing access to multi-million and multi-billion-word corpora of present-day and historical English” (Davies, 2019), moreover it can serve as a repository of over 500 corpora across 95 languages (Kilgarriff et al., 2014). Computerized corpora have proved to be excellent recourses for a wide range of research tasks connected with learning the language (Andrushenko, 2021) since they facilitate the automated search of the linguistic data, assist in analysing language phenomena based on significantly large collections of texts that represent various natural languages (Davies, 2019; Johansson, 2009; McEnery&Hardie, 2015; Rissanen, 2009). Modern linguistics has been continually and constantly enriched with the new collective monographs (see.: López-Couso et al., 2016; Whitt, 2018), manuals (see: Collins, 2019; Lange&Leuckert, 2020; Stefanowitsch, 2020) and articles (Andrushenko, 2022; Anokhina, 2023; Lavidas&Haugh, 2020) that represent fundamental theoretical and methodological ground for research and specify the possibilities of different software aimed at corpora investigation. Despite simplifying the data search on the one hand, the corpus system requires knowledge of different approaches and methodologies of investigation, on the other hand. This presupposes competence in statistic verification that helps to support or disprove the hypothesis made (Andrushenko, 2021).

Undoubtedly, artificial intelligence programs are powerful tools for an automated analysis of linguistic phenomena, among which LancsBox stands out as a new generation software package for the study of languages. Initially developed at the University of Lancaster in 2015 (Brezina et al., 2015), it can work with the existing corpora, that have recently been elaborated, or with linguist’s own data assisting in visualizing language facts, which presupposes their automatic annotation for part-of-speech. The major features of the software are 1) working with user’s data or existing corpora that can be loaded in various formats (pdf, xml, docx, .doc, etc.); 2) language facts visualization; 3) analyzing the data irrespective of the language; 4) automatic annotation of data for part-of-speech, 5) compatibility with the main operating systems (Mac, Windows, Linux) (Brezina et al., 2018). The main asset of the software, according to its principal developers, lies in “automated research on word associations, identifying collocates based on traditional three criteria: distance (specifying the span around a node word, ‘collocation window’), frequency (an important indicator of typicality of word association) and exclusivity” (Brezina, 2018). The other criteria that should be taken into account are directionality (which presupposes to measure the attraction strength between collocates), dispersion (the distribution of the node

and the two adjacent words in text corpora) and type-token distribution among collocates (viz. the strength of the collocational relationship and the level of competition for the slots around the node word from other collocate type) (Gries, 2013). Additionally, the developers of LancsBox take into account the connectivity between individual collocates (Brezina et al., 2015). Apart from working with user's data, LancsBox grants access to built-in corpora that approximately include 1,000,000 tokens each. Such corpora can be exemplified by American and British English text samples (AmE06; BE06, BNC1940-baby, etc.) (Brezina et al., 2020). Non-European languages are brought forward by Lancaster Corpus of Mandarin Chinese (L-C-M-C), etc. The full list of corpora accessible for download is given in Figure 1.

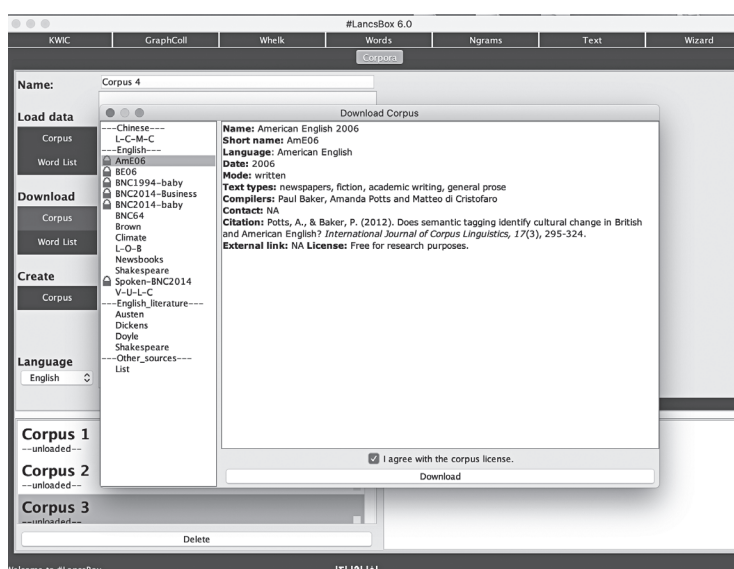


Fig. 1. The list of available built-in corpora in LancsBox

Methodology. The two corpora selected for the pilot investigation and licensed by LancsBox software are AmE06 (American English) and BE06 (British English) representing Brown Family of corpora (Baker, 2009). These are a “carefully balanced set of samples with approximately the same number of words (1,000,000+) for each genre coming from a single period of time” (Potts & Baker, 2012), i.e. the year of 2006. This allows comparing words within the same time frame and different types of English (in case of the current study the usage of the lexeme *European* has been estimated). Each sample from different genres in corpora amounts to over 2,000 words. The allotment of samples per genre is as follows: press editorials (27), press reportage (44), press reviews (17), skills, biographies and essays (75), trades and hobbies (36), religion (17), popular lore (48), miscellaneous (reports, science (academic prose) (80), official documents (30), mystery and detective fiction (24), general fiction (29), western and adventure fiction (29), romantic fiction (29), science fiction (6), humor (9) (Lawrence, 2019).

To simplify the data search and visualize the results obtained the following tools from LancsBox package have been used: KWIC (enables co-textual information about the token under scrutiny. It generates a list of all instances of a search term in a corpus in the form of a concordance (Andrushenko, 2023). Double clicking on the node opens a pop-up window with a larger number of the texts which allows investigating the word in a broader

context), Words (which main function is to seek words belonging to the same word class), GraphColl (provides data on the collocational patterning of the node search. It can visualize both right and left collocates simultaneously or separately depending on the parameters identified for a collocation network graph taking into account three parameters: strength, frequency, position) (Brezina & Porizka, 2021). The Words tool provides in-depth analysis of frequencies of types, part-of-speech categories and lemmas as well as allows to compare corpora using the keywords technique. The Ngrams tool enables the analysis of frequencies of different ngram types, lemmas and part-of-speech categories and it also facilitates the comparison of corpora using the key ngram technique. (Brezina, 2018; Brezina et al., 2020).

Results and discussion. The statistical analysis bar shows that the word *European* in BE06 corpus occurs 175 times (1.76 per 10K) in 72 texts out of 500, while the frequency of the same word in Am06 is significantly lower, viz. 96 occurrences (0.96 per 10K) in 53 out of 500 texts, which can be explained by the cultural differences of speakers in terms of their interest to the current events. The comparative data for both language variations are presented in Figure 2.

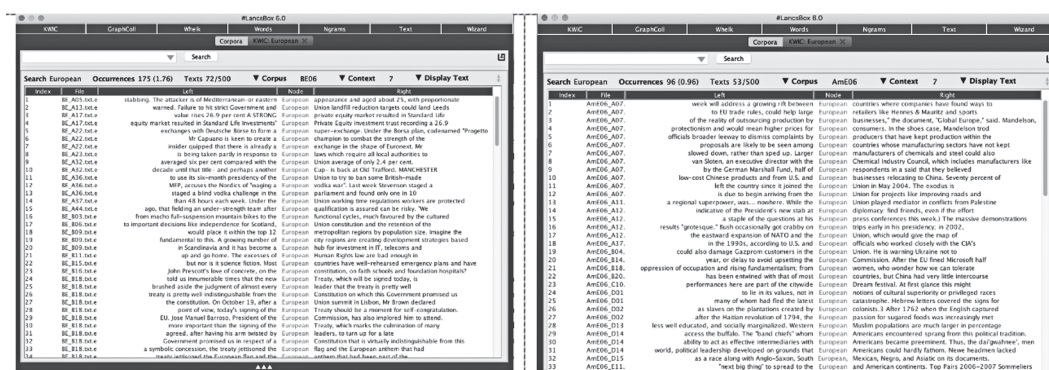


Fig. 2. Frequency of search term *European* in BE06 and Am06 (LancsBox)

The GraphColl tool has allowed singling out collocates of *European* in Am06 and BE06 using the Collocation frequency (01 – Freq (5.0), L5-R5, C: 5.0 – NC: 5.0) (Brezina, 2018). Word associations of the top 10 collocates in both corpora are presented in Tables 1-2.

Table 1

Collocates of search term *European* in AmE06

ID	Position	Collocate	Stat (Freq)	Freq coll	Freq corpus
1	L	the	56	56	59942
2	L	of	34	34	30270
3	L	and	28	28	28797
4	L	in	25	25	19813
5	L	a	24	24	23381
6	L	to	20	20	25899
7	L	with	11	11	6961
8	L	for	10	10	8884
9	L	on	9	9	6866
10	L	that	8	8	11842

Table 2

Collocates of the search term *European* in BE06

ID	Position	Collocate	Stat (Freq)	Freq coll	Freq corpus
1	L	the	169	169	58919
2	L	of	76	76	30653
3	R	and	48	48	27911
4	L	to	44	44	26189
5	L	a	37	37	22758
6	R	in	33	33	19264
7	R	union	31	31	101
8	L	for	19	19	9252
9	R	on	17	17	7382
10	M	that	16	16	10231

The comparison of the search term in two Corpora has shown that the frequencies of the first three collocates are almost identical. Hence, the word *European* is most often used with the article *the*, preposition *of* and conjunction *and*. However, there is a slight difference in right and left dislocation of collocates when it comes to the conjunction usage in both Corpora. Moreover, the lexeme *European* rather frequently collocates with the noun *union* (31 collocates out of 175 amounting to 17.71%) in British English and further investigation of the American English (AmE06) has shown that it occupies the 15th place in terms of frequency being represented by 7 collocates only (7.29%). The collocation networks for both variations of the language are given in Figures 3–4.

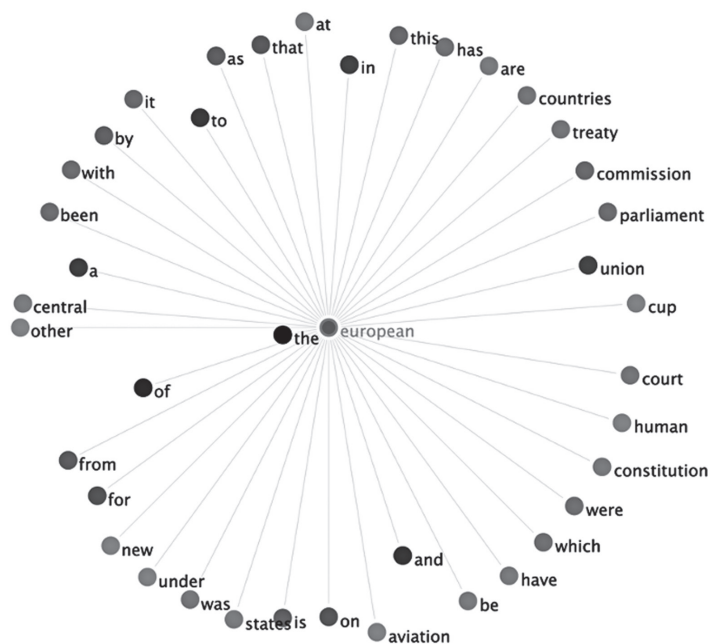


Fig. 3. Collocation network: *European* in BE06

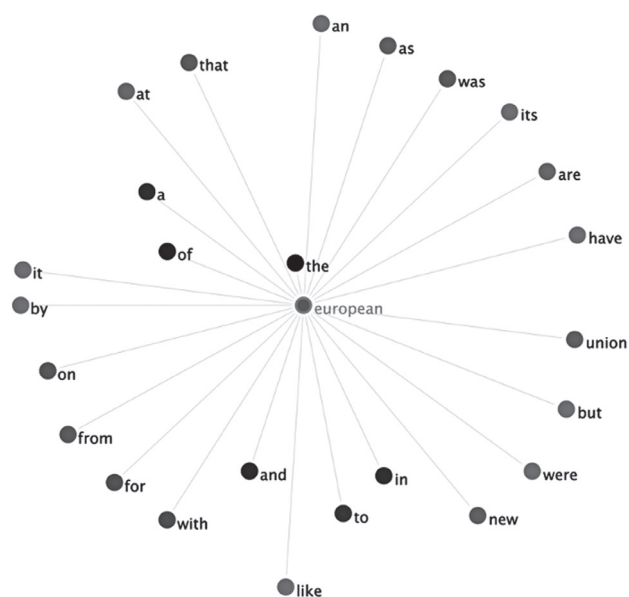


Fig. 4. Collocation network: *European* in AmE06

The collocation networks in Figures 3 and 4 indicate that the most typical collocates in British English are exemplified by N+N structures: *European Parliament*, *European Countries*, *European Court*, *European Treaty*, *European Commission*, *European Union*. This can suggest that the European studies are given a significant coverage in this language variation, while in American English the most frequent tokens with N+N are found in the single collocate *European Union*.

The t-test ($t(782.19) = -2.16, p = 0.031$) has revealed a statistically significant difference between the corpora with regard to the linguistic variable *European*. This result is visualised in Fig. 5 below. Cohen's d ($-0.14, 95\% \text{ CI } [-0.26, -0.01]$) showed a minimum effect. Figure 5 shows error bars plot in both corpora.

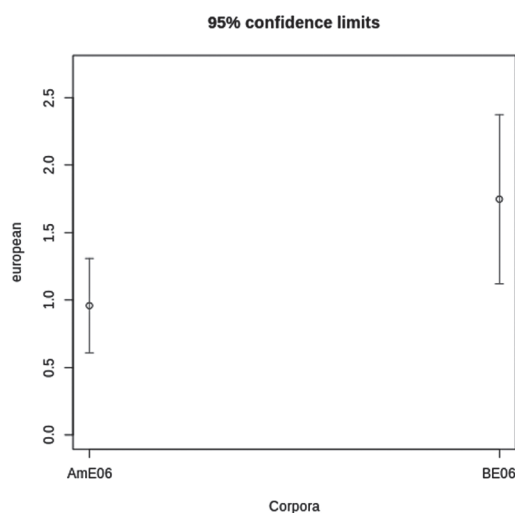


Fig. 5. Error bars plot for *European* in AmE06 and BE06

LancsBox also enables to trace the frequency of the lexical unit in the corpus. Thus, such software tool as Words visualizes the most frequent words in the selected corpus, which is illustrated in Figure 6 based on AmE06.

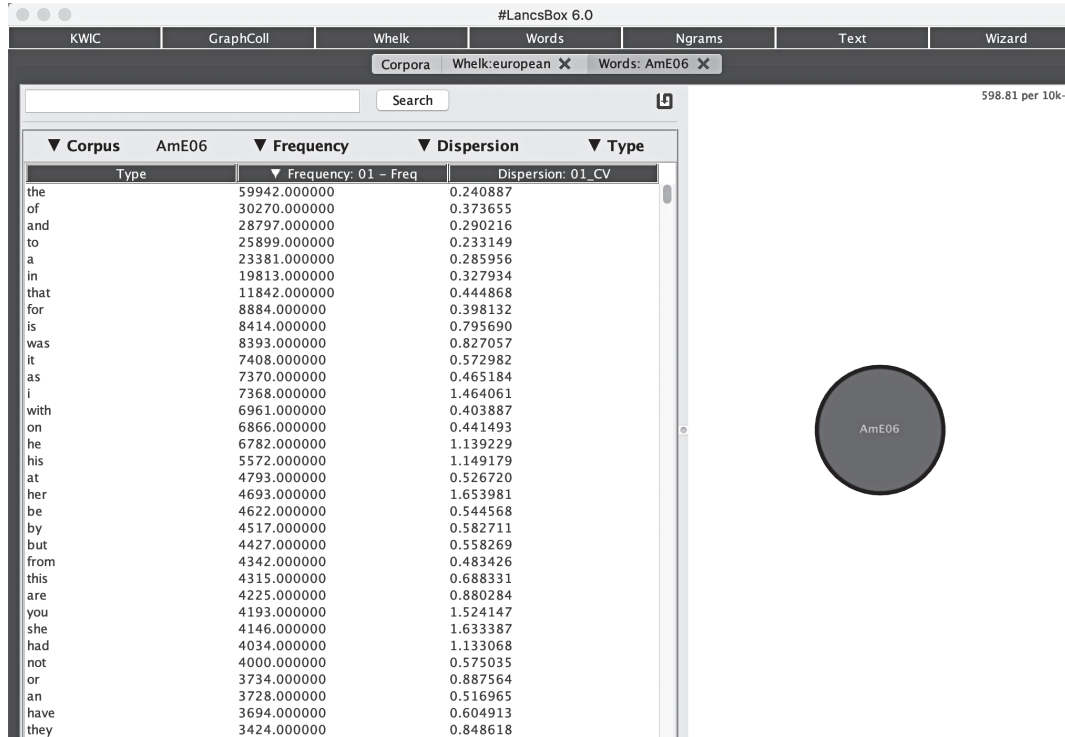


Fig. 6. The most frequent words in AmE06

The study of the word *European* in AmE06 has indicated that it belongs to non-frequent vocabulary with the dispersion that amounts to 4.147510 (Figure 7). The same is true for BE06 corpus (Figure 8), where the collocate *the European* has a bit higher dispersion of 5.239762.

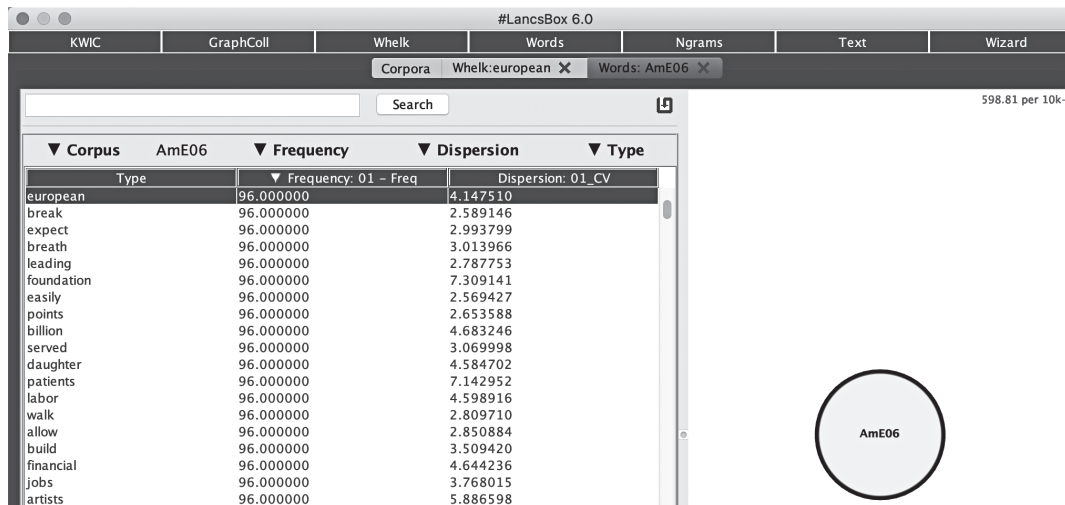


Fig. 7. The frequency of the lexical unit *European* in AmE06

Corpus	BE06	Frequency	Dispersion	Type	Grams
		▼ Frequency: 01 – Freq	Dispersion: 01_CV		
the european		80.000000	5.239762		
that of		80.000000	2.848186		
the royal		80.000000	4.813126		
then the		80.000000	2.732084		
her and		80.000000	3.284921		
don't know		80.000000	3.004473		
at his		80.000000	2.782502		
was going		80.000000	2.801906		
until the		80.000000	3.177432		
then he		80.000000	3.788755		
idea of		80.000000	2.871174		
to meet		79.000000	2.966117		
the department		79.000000	5.103759		
although the		79.000000	3.055424		
the current		79.000000	3.506226		
them and		79.000000	2.569550		
get a		79.000000	2.931740		
had no		79.000000	2.941656		
compared with		79.000000	4.539190		
to put		79.000000	2.678127		
it wasn't		79.000000	3.240828		
you think		78.000000	3.148725		
the		78.000000	3.331919		
make the		78.000000	2.728998		
among the		78.000000	2.645804		
i thought		78.000000	3.166915		
with its		78.000000	2.833674		
will not		78.000000	2.966579		
to ensure		78.000000	3.066056		
she would		78.000000	4.094865		
in and		78.000000	2.904245		
the us		78.000000	3.668404		
difficult to		78.000000	2.783404		
in addition		78.000000	2.919151		

Fig. 8. The frequency of the lexical unit *European* in BEE06

Concluding remarks. The automated analysis of the word *European* has shown that owing to LancsBox software the lexical unit can be analysed in parallel in two different corpora: BE06 and AmE06. The software tools allow visualizing not only the most frequent collocates with the word based on left, middle and right dislocation but also enable to find the most regular collocation patterns for every language variation. As the investigation has indicated, N+N structure is frequently traced in BE06, although this tendency is not characteristic of AmE06. Moreover, the LancsBox provides the opportunity to trace the frequency of the word usage in both corpora representing and visualizing the peculiarities of both language variations related to cultural differences. This software can be implemented for analysis of user's corpora that will further compile the Multilingual language corpus for European studies.

REFERENCES

- Andrushenko, O. (2021). Information-structural transformations of additive adverb EVEN (a case study of the English language written records and corpora of the XII–XVII c.). *Messenger of Kyiv National Linguistic University. Series Philology*. Volume 24, No. 1, pp. 16–32. DOI: 10.32589/2311-0821.24%20(1).2021.236109.
- Andrushenko, O. (2022). The Scope of *just*: evidence from information-structure annotation in diachronic English Corpora. In N. Sharonova, V. Lytvyn, et al. (Eds.), *Proceedings of the 6th international conference on computational linguistics and intelligent systems (COLINS 2022), Vol. I: Main Conference, Gliwice, Poland, May 12–13, 2022* (pp. 677–696). Available online: <https://ceur-ws.org/Vol-3171/paper51.pdf>
- Andrushenko, O. (2023). Particularizing focus markers in Old English: just the case of adverb polysemy? *Lege Artis: Language yesterday, today, tomorrow*. (Accepted for publication, date of publication: December 2023).

- Anokhina, T. (2023). Newspaper subcorpus (subcorpus of the modern european media) in the structure of the multilingual corpus. *Philological Treatises*. Volume 15, No. 1, pp. 7–15. DOI: 10.21272/Ftrk.2023.15(1)-1.
- Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics*, 14 (3), 312–337. DOI: 10.1075/ijcl.14.3.02.bak.
- Baker, P. (2010). Corpus methods in linguistics. In L. Litosseliti (Ed.), *Research methods in linguistics* (pp. 93–113). London, New York: Continuum.
- Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.
- Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20 (2), 139–173.
- Brezina, V., Porizka, P. (2021). Kolokační grafy a sítě s použitím nástroje #LancsBox: aplikace v angličtině a češtině. *Časopis pro moderní filologii*, 103, Č. 1, 36–59. DOI: 10.14712/23366591.2021.1.
- Brezina, V., Timperley, M., & McEnery, T. (2018). #LancsBox 4.x [software]. Available online: <http://corpora.lancs.ac.uk/lancsbox>.
- Brezina, V., Weill-Tessier, P., & McEnery, T. (2020). #LancsBox 5.x and 6.x [software]. Available online: <http://corpora.lancs.ac.uk/lancsbox>.
- Collins, L. (2019). *Corpus linguistics for online communication: A guide for research*. New-York: Routledge.
- Davies, M. (2019). The best of both worlds: Multi-billion word “dynamic” corpora. In P. Banský et al. (Eds.), *Proceeding of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019* (pp. 23–28). Manheim: Leibniz Institute für Deutsche Sprache. DOI: 10.14618/ids.pub.8998.
- Gries, S. (2013). 50-something years of work on collocations: What is or should be next... *International Journal of Corpus Linguistics*, 18 (1), 137–166. DOI: 10.1075/ijcl.18.1.09gri.
- Johansson, S. (2009). Some aspects of the development of corpus linguistics in the 1970s and 1980s. In Anke Lüdeling & Merja Kytö (Eds), *Corpus linguistics: An international handbook* (pp. 33–53). Berlin: De Gruyter.
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubiček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1 (1), 7–36. DOI: 10.1007/s40607-014-0009-9.
- Lange, C. & Leuckert, S. (2020) *Corpus linguistics for world Englishes: A guide for research*. New-York: Routledge.
- Lavidas, N. & Haugh, D.T.T. (2020). Postclassical Greek and treebanks for a diachronic analysis. In D. Rafiyenko & I. Seržant (Eds.), *Postclassical Greek: contemporary approaches to philology and linguistics* (pp. 163–202). Berlin: Walter de Gruyter.
- Lawrence, S. (2019). *A rite of the edge: The language of baptism and christening in the church of England*. London: SCM Press.
- López-Couso, M. J., Méndez-Naya, A., Núñez-Pertejo, B. P., & Palacios-Martínez, I. M. (2016). *Corpus linguistics on the move. Exploring and understanding English through corpora*. Leiden, Boston: Brill Rodopi.
- McEnery, T., & Hardie, A. (2015) *Corpus Linguistics*. New-York: Routledge.
- O’Keeffe, A., McCarthy, M. (2021). *The Routledge Handbook of Corpus Linguistics*. New-York: Routledge.
- Potts, A., & Baker, P. (2012). Does semantic tagging identify cultural change in British and American English? *International Journal of Corpus Linguistics*, 17 (3), 295–324.
- Rissanen, M. (2009). Corpus linguistics and historical linguistics. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 53–68). Berlin: De Gruyter.

Stefanowitsch, A. (2020). *Corpus linguistics: a guide to the methodology*. Berlin: Language Science.


Whitt, R. (2018). Using diachronic corpora to understand the connection between genre and language change. In R. Whitt (Ed.), *Diachronic corpora, genre, and language change*, (pp. 1–18), Amsterdam, Philadelphia: John Benjamins Publ.

Co-funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Дата надходження до редакції 05.04.2023

Ухвалено до друку 22.06.2023

Відомості про автора

<p>Андрушенко Олена Юріївна,</p> <p>кандидат філологічних наук, доцент, докторантка кафедри германської і фіно-угорської філології імені професора Г. Г. Почепцова Київського національного лінгвістичного університету</p> <p>e-mail: and olenka@gmail.com</p>		<p>Сфера наукових інтересів:</p> <p>історичний синтаксис англійської мови, інформаційна структура речення, корпусна лінгвістика</p>
--	---	--