UDC 81’246+81’33]:001(4)(045)
DOI https://doi.org/10.26661/2414-1135-2025-99-11


# MULTILINGUAL CORPUS AS A RESOURCE FOR EUROPEAN STUDIES RESEARCH

**Korolyova A. V.**
*Vice-Rector for Research and International Cooperation*
*Kyiv National Linguistic University*
*Velyka Vasylkivska str., 73, Kyiv, Ukraine*
*orcid.org/0000-0001-5541-5914*
*alla.korolyova@knlu.edu.ua*

***Key words:*** *multilingual text corpus, European studies, corpus analysis methodology, linguistic information retrieval system.*

The article discusses general and specialised text corpora and their resources for European studies. The main focus is on the development and presentation of a methodology for creating a specialised multilingual text corpus for the analysis of narrow EU topics (political, diplomatic, legal, economic, academic, etc.), authored by Professor V. Zhukovska. The stages and procedures of building a specialized multilingual text corpus for researching various EU domains are described and characterized, encompassing the following research algorithms: corpus design; identification of target topics and queries; determination of corpus type, size, and sampling strategies; identification of sources and data collection; data input; conversion and graphematic analysis of selected texts; text tagging/annotation; correction of automated tagging/annotation results; conversion of annotated texts into the structure of a specialized linguistic information retrieval system; and providing access to the corpus. The article outlines further research prospects with the specialized European studies corpus, which enables corpus-based linguistic analysis of the lexico-semantic features of European legal language (terminology systems, polysemy, homonymy of legal terms), syntactic patterns and typical grammatical constructions, as well as rhetorical strategies and argumentation models used in EU academic legal discourse. The corpus can also be applied to the compilation of specialized dictionaries (explanatory, bilingual, terminological databases) of European law and to the creation of glossaries of collocations and phraseological units for legal practitioners and translators.

---

# МУЛЬТИЛІНГВАЛЬНИЙ КОРПУС ЯК РЕСУРС ДЛЯ ДОСЛІДЖЕННЯ ЄВРОПЕЙСЬКИХ СТУДІЙ

**Корольова А. В.**
*проректор з наукової роботи і міжнародного співробітництва*
*Київський національний лінгвістичний університет*
*вул. Велика Васильківська, 73, Київ, Україна*
*orcid.org/0000-0001-5541-5914*
*alla.korolyova@knlu.edu.ua*

***Ключові слова:***
*мультилінгвальний корпус текстів, європеїстика, методика корпусного аналізу, лінгвістична інформаційно-пошукова система корпусу.*

У статті розглянуто загальномовні і спеціалізовані корпуси текстів та їхні ресурси для дослідження європеїстики. Основну увагу приділено розробці й презентації методики створення спеціалізованого мультилінгвального корпусу текстів для аналізу вузької тематики ЄС (політичної, дипломатичної, правової, економічної, академічної тощо), автором якої є професор В. Жуковська. Описано і схарактеризовано етапи і процедури

створення спеціалізованого мультилінгвального корпусу текстів для дослідження різних сфер діяльності ЄС, які включають такі алгоритми дослідницьких дій, як: проєктування корпусу, окреслення цільової тематики і запитів, визначення типу спеціалізованого мультилінгвального корпусу текстів, його обсягу та стратегій вибірки, визначення джерел мультилінгвального матеріалу та збір даних для корпусу, введення даних корпусу, конвертування й графематичний аналіз відібраних текстів, розмітка / анотація текстів корпусу, коригування результатів автоматичної розмітки / анотації, конвертування розмічених текстів у структуру спеціалізованої лінгвістичної інформаційно-пошукової системи, забезпечення доступу до корпусу. Окреслено перспективи подальшої роботи з уже створеним спеціалізованим корпусом текстів для дослідження європеїстики, який дає можливості виконувати корпусно-лінгвістичний аналіз лексико-семантичних характеристик мови європейського права (терміносистеми, полісемії, омонімії правових термінів), синтаксичних патернів та типових граматичних конструкцій, риторичних стратегій та аргументаційних моделей, що використовуються в академічному правовому дискурсі ЄС. Також ресурси створеного корпусу можна використати для укладання спеціалізованих словників (тлумачних, перекладних, термінологічних баз даних) з європейського права, для створення глосаріїв колокацій та фразем європейського права для юристів та перекладачів.

**The statement of the problem.** Corpus linguistics explores the concept of a corpus of texts, which, among other interpretations, is understood as a collection of language fragments selected according to clear linguistic criteria for use as a specific language model [McEnery & Wilson, 2001, p. 29].

The development of computer technologies has resulted in the creation of various types of text corpora, which can satisfy specific research needs. Closely related to the concept of a text corpus are the concepts of concordance and corpus manager or corpus manager [Zhukovskaya, 2015, 2020, 2023; Korolyova, 2023, 2024]. Searching for data in a text corpus allows you to build a concordance for any word, which is a specialised search engine that includes software tools for searching data in the corpus, obtaining statistical information and providing results in a user-friendly form [O'Keeffe et al., 2007]. It should be added that when using web space as a text corpus, search engines such as AntConc, Sketch-Engine can perform the role of corpus managers.

Recently, specialised text corpora have become the most popular corpus resources among users. These include the Multilingual Text Corpus for European Studies developed by KNLU team https://mcresr.knlu.edu.ua/corps/, which is an additional resource of scientific and media information dedicated to covering the activities of all European institutions, as a supplement to general language (monolingual and multilingual) text corpora and ready-made specialised text corpora on EU topics [Korolyova, 2023, 2024].

V. Zhukovska has her own way of working with corpus resources and tools for studying texts on European studies, which has three areas of analysis: 1) analysing how different areas of EU activity are represented in general language text corpora, i.e. in a variety of linguistic environments (different genres, years, etc.); 2) analysis of thematic contexts (legal, political, environmental, etc.) of European studies in ready-made specialised text corpora; 3) analysis of narrow topics of European studies in a specialised text corpus created specifically for this task [Zhukovska, 2018, 2020]. This methodology was successfully implemented during the training course 'Digital Tools for European Studies Research' at KNU (https://mcresr.knlu.edu.ua/trainings/).

**Purpose and Objectives.** The aim of this article is to demonstrate the advantages of a specialized multilingual corpus as an optimal resource for European studies.

**Object of study:** general-purpose and specialized text corpora (in English and Ukrainian) for EU-related research.

**Subject of study:** the process of creating a multilingual corpus for narrow-topic European studies analysis.

**Main Research Findings**

Corpus-based analysis of broad EU discourse focuses on structuring thematic areas represented in texts of various styles. To achieve this, V. Zhukovska developed a system of lexical queries covering major conceptual-semantic domains verbalized in EU-related texts (https://mcresr.knlu.edu.ua/trainings/).

Sample query domains include:

1. General concepts (European Union, EU integration, European Community).

2. Politics and governance (EU Institutional Politics, European Commission, Council of the EU,

European Council, Enlargement and Membership, EU enlargement, accession to the EU, governance models, economic integration, trade policy).

3. Socio-cultural dimension (EU foreign policy, external relations, Common Foreign and Security Policy, strategic partnerships).

4. Specific policy areas (environment, climate policy, European Green Deal, digital policy, agriculture, fisheries).

**1. The analysis of the representation of various areas of EU activity can be initiated with general language text corpora (in English and Ukrainian).**

**General language text corpora:**

The *Corpus of Contemporary American English* (COCA) covers the period from 1990 to 2019 and is hosted on the platform *english.corpora.org*. Using the *List* and *Chart* options, one can analyse the usage of key concepts such as *integration*, *European Union*, *European*, *Europe*, firstly, across the entire corpus and trace changes in the contexts of their functioning over the years, and secondly, limit the search by styles and genres, for instance, mass media (multimedia).

*English Web 2021* (enTenTen21) on the *Sketch Engine* platform belongs to the related TenTen corpora family, compiled in over 40 languages with a total volume of 52 billion words. The texts in these corpora are annotated according to genre affiliation and thematic domain. Based on the corpus data in *English Web 2021* (enTenTen21), it is also possible to analyse key thematic units such as *Europe*, *Ukraine*, *integrity*, *EU integration*, among others.

**2. It is advisable to proceed with the available specialised corpus resources:**

Corpora of European-themed texts:

*European Parliament Proceedings Parallel Corpus 1996–2011* (*Europarl Parallel Corpus*)

*Paralela*

*ParaCrawl Corpus*

*DCEP: Digital Corpus of the European Parliament*

*Europarl spoken parallel corpus*

*European Parliament Translation and Interpreting Corpus*

*Legal Corpora*

*Parliamentary Corpora* (including *ParlaMint: Comparable and Interoperable Parliamentary Corpora*, which comprises 17 multilingual comparable corpora of parliamentary debates)

Corpora of mass media texts applicable for analysing specific areas of EU activity:

*Mass media corpora on the Sketch Engine platform:*

– English Trends corpus

– Ukrainian Trends: a daily-updated monitor corpus of news articles

– ukTenTen: Ukrainian corpus from the Web

– EUR-Lex parallel corpus

– DOAJ corpora – Directory of Open Access Journals

– Brexit corpus (English)

*Mass media corpora on the CLARIN platform:*

– Newspaper Corpora

– Parliamentary Corpora

– UKParl Dataset (2014–2013, 2014–2015, 2015–2016)

*Other resources for analysing the broader EU discourse:*

– Eurotermbank

– DGT-Translation Memory

– Unleashing European Patent Translations

**3. Finally, it is possible to create a customised specialised multilingual text corpus for researching various spheres of EU activity.**

The algorithm for building such a corpus consists of the following sequential procedures:

**3.1. Corpus design** begins with formulating the research question and defining the corpus structure. The specification of the target thematic scope and queries depends on the dimensions of EU studies, which may encompass legal, political, economic, academic, sociocultural, and historical aspects, as well as international relations.

**3.2. The process of corpus creation** requires the following key stages:

1. Determining the type of specialised multilingual text corpus, its size, and sampling strategies.

2. Identifying the sources of multilingual material and collecting data for the corpus.

3. Inputting the corpus data.

4. Converting and performing graphematic analysis of the selected texts.

5. Tagging/annotating the corpus texts.

6. Adjusting the results of automatic tagging/annotation.

7. Converting the annotated texts into the structure of a specialised linguistic information retrieval system.

8. Ensuring access to the corpus.

**Determining the type and size of the corpus.** In general, there are two main types of corpora:

1. *Static corpus* – contains materials for a defined period (e.g., articles from certain years) to capture the state of a topic at a given time slice, such as the evolution of EU discourse before/after Brexit.

2. *Monitor corpus* – regularly updated (monthly/annually) with new materials to track changes over time, for instance, the dynamics of EU political discourse.

The approximate size of the corpus can be small (50–100 texts), medium (100–300 texts), or large (up to 1.000 texts, which requires semi-automatic search and data collection). Corpus sampling strategies should ensure a balance of genres and temporal coverage. For balanced thematic representation of texts in a

corpus (e.g., for 100 texts), V. Zhukovska recommends a specific percentage distribution of topics.

**Identifying sources of multilingual material and collecting data for the corpus.** The key stage in creating a specialised corpus is selecting and systematising sources. The sources for the corpus may be either openly available or have restricted/limited access.

Table 1

**Balanced Thematic Representation of Texts in the Corpus (N = 100)**

| Theme | % of Corpus | Number of Texts |
|---|---|---|
| General EU topics | 10% | 10 |
| Politics and Governance | 20% | 20 |
| Law and Legislation | 15% | 15 |
| Economics and Trade | 15% | 15 |
| Sociocultural Issues | 10% | 10 |
| Foreign Affairs | 15% | 15 |
| Specific Policies (e.g., Climate) | 10% | 10 |
| Critical Events (e.g., Brexit) | 5% | 5 |

**It should be noted** that if the corpus is made publicly available online, copyright issues regarding full-text materials must be carefully considered. The use of copyrighted materials, such as textbooks and monographs, requires the written consent of rights holders, which may significantly slow down the compilation of such a corpus. Recommended sources include open-access scholarly journals on EU-related topics, analytical reports and review documents (e.g., European Commission Reports, European Parliament Briefings), legal and regulatory documents, materials from specialised scientific conferences, as well as specialised platforms and websites.

Google Scholar, Semantic Scholar, Research Gate, JSTOR, and DOAJ (Directory of Open Access Journals) were used to search for scientific journals and articles when compiling the multilingual corpus. Specialised platforms and websites included Eurodocs, the European Commission, and the Council of Europe.

When the data collection is done, the materials get some technical prep. During the data entry stage, the selected texts are organised, given proper names, and annotated. The file structure is set up with separate folders for original files (PDF, HTML) and converted files (TXT format, assigning each file a unique name, which will be displayed in the corpus as a document (text) identifier, compiling a table of correspondence between the file name in the corpus and its full bibliographic description.

For effective corpus work using a corpus manager (e.g., *AntConc*), the selected texts must be prepared. **Text preparation for analysis** involves conversion and graphematic processing. This stage includes re-encoding, removing or transforming non-verbal elements (images, tables, graphs, formulas), removing hyphenation, and deleting structural elements. To convert files (e.g., PDF to TXT), standardise encoding (UTF-8), and perform related tasks, applications from the *AntLab* platform are used – in particular, *AntFileConverter* for converting PDF files to TXT format and *EncodeAnt* for encoding texts in the UTF-8 standard.

**Corpus annotation** (if required). If the research involves grammatical or syntactic analysis, the texts should be annotated – for instance, part-of-speech tagging using *TagAnt*.

After cleaning and annotation, the corpus is uploaded into a **specialised linguistic information retrieval system**, such as *AntConc*.

**Linguistic Analysis of the Corpus Using AntConc**

*AntConc* is one of the most efficient free tools for corpus analysis. Its functionality enables a comprehensive multi-aspect quantitative and qualitative analysis, including:

**Lexical analysis,** which primarily involves examining frequency vocabulary, terminology, and thematic lexicon.

**Frequency list:** Using the *Word List* function, one can identify high-frequency vocabulary characteristic of EU discourse. For example, in the legal subcorpus of EU texts, typical high-frequency items include *law*, *EU*, *Court*, *Member State*, *directive*, *regulation*, *principle*, *decision*, *treaty*, *judgment*, *legal*, *rights*, *Union*, *national*, *international*, *policy* – forming the thematic core. Terms such as *jurisdiction*, *precedent*, *implementation*, *harmonisation*, *infringement*, *enforcement*, *liability*, *supremacy*, *proportionality* constitute the terminological core and serve as indicators of this subcorpus.

**Keywords:** Using the *Keyword List* function in comparison with a reference corpus (e.g., *Brown Corpus*) allows identification of keywords that form the EU law terminological core (e.g., *EU*, *Union*, *directive*, *regulation*, *treaty*, *court*, *jurisdiction*, *harmonisation*, *rights*, *order*, *supremacy*, *proportionality*, *subsidiarity*). Additionally, units denoting key EU legal principles and actors (e.g., *Commission*, *Parliament*, *justice*, *integration*, *citizen*, *internal market*, *competition*) are revealed.

**Collocational and colligational analysis:** This analysis identifies typical combinations of selected units, their "environment", and stable expressions that form their collocational profiles. Using the *Collocates*, *Clusters*, and *N-grams* functions, one can identify frequent word combinations (*national law*, *legal system*) and explore the meaning of terms in context, comparing collocational profiles of semantically related terms (*directive* vs *regulation*). The *N-grams*

tool also detects multiword units (2-, 3-, 4-, 5-word sequences).

**Distribution visualisation:** Functions such as *Plot*, *KWIC*, and *File View* help determine whether the distribution of a given item is uniform across the corpus or concentrated in specific text segments, indicating thematic clusters (e.g., sections of texts focused on a specific topic).

**Grammatical analysis** focuses on grammatical patterns and syntactic constructions found in thematic subcorpora. This includes examining typical syntactic models (e.g., the use of passive constructions, complex sentences with specific conjunctions, and modal constructions expressing normativity or obligation using verbs and nouns with modal meaning such as *must*, *should*, *obliged*, *necessary*, *duty*).

The use of **morphosyntactic analysis** (when applying the POS-tagger *TagAnt* and having POS-tagged data) enables searching by part-of-speech tags (e.g., *_NN*, *_VB*, *_JJ*) to determine their frequency, and by tag combinations to examine typical grammatical structures (e.g., adjective-noun groups *JJ _NN*).

**Discourse-semantic analysis** includes examining evaluative connotation, allowing the identification of "positive" and "negative" connotations through typical collocability. For example, analysis of frequent nouns with premodifying adjectives shows that *effectiveness* acquires a positive connotation when combined with *increased*, *institutional*, *enhanced*.

**Frame analysis**, based on the construction of semantic frames (as in *FrameNet*) and used for thematic profiling of discourse (term by V. Zhukovska), enables the identification of typical themes and subthemes and the modelling of typical semantic scenarios reflecting the conceptual structure of each subcorpus. For example, in analysing the key concept *CJEU* (*Court of Justice of the European Union*), frequent lexemes collocating with *court* (*effect*, *impact*, *development*, *evolution*, *shaping*, *transformation*, *consolidation*, *establishment*, *clarity*, *consistency*, *uniformity*, *integration*) describe outcomes, consequences, or influence of Court decisions.

Based on lexemes and their meanings obtained from the corpus, the data are mapped to the corpus-based lexicographic resource *FrameNet* to identify potential frames representing relevant situations in the subcorpus. Examples of potentially relevant *FrameNet* frames include:

**Causation:** Describes situations where one event or state causes another (*effect*, *impact*, *transformation*). *Frame Elements (FE)*: Cause, Effect, State_of_affairs.

**Change:** Describes the process in which an entity moves from one state to another (*development*, *evolution*, *transition*).

FE: Entity, Initial_state, Final_state, Manner, Degree.

**Coming_to_be / Creation:** Describes the process by which something comes into existence or is created (*shaping*, *creation*).

FE: Creator, Created_entity, Means.

**Achieving_position:** Describes attaining a certain position or status (*establishment*, *recognition*).

FE: Agent, Theme, Result_position.

**Impact:** Describes the influence of one entity on another (*impact*, *clarity*, *integration*).

FE: Impactor, Impactee, Impact_type.

**Conclusions and prospects for further developments in this area.**

In concluding, we note that the most optimal resource for studying European studies is the created 'Multilingual Corpus of Texts,' which consists of subcorpora of texts on narrow topics devoted to revealing various areas of EU activity (legal, political, diplomatic, academic, etc.).

The specialized corpus of texts and the proposed methods of corpus analysis open up wide opportunities for further research into the lexical and semantic characteristics of European legal language (terminology, polysemy, homonymy of legal terms), syntactic patterns and typical grammatical constructions, rhetorical strategies and argumentation models used in EU academic legal discourse, etc.; for compiling specialised dictionaries (explanatory, translation, terminological databases) on European law, for creating glossaries of collocations and phrases of European law for lawyers and translators.

*This project is co-funded by the European Union. The views expressed, however, are those of the author alone and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency. Neither the European Union nor the funding body is responsible for the views expressed in this article.*

## BIBLIOGRAPHY

1. Жуковська В.В. Корпусна лінгвістика: історія становлення та сучасний стан. / А. В. Сингаївська (Ред.). Актуальні лінгвістичні студії : навчальний посібник. 2015, с. 168–204. Житомир : Вид-во ЖДУ імені Івана Франка.
2. Жуковська В.В. Застосування корпусних технологій у навчанні та вивченні іноземної мов. *Матеріали онлайн-конференції «Актуальні проблеми сучасної лінгвістики та методики викладання мови і літератури»*, м. Житомир, 7–11 лютого, 2018 р., с. 39–50. Житомирський державний університет імені Івана Франка. 2018.URL:https://nniif.org.ua/File/18zvvzkt.pdf.
3. Жуковська В.В. Лінгвістичний корпус як новітній інформаційно-дослідницький інструментарій сучасного мовознавства. *Вчені*

*записки Таврійського національного університету ім. В.І. Вернадського. Серія : Філологія. Соціальні комунікації*, 2020. *31*(70), № 3, ч. 1, 113–119. https://doi.org/10.32838/2663-6069/2020.3-1/20.

4. Жуковська В.В. Корпусні технології та жанрово-аналітичний підхід у навчанні англійської мови для академічних цілей. *Дискурс професійної і творчої комунікації: лінгвокультурний, когнітивний, перекладацький та методичний аспекти* : збірник матеріалів VIII Міжнародної науково-практичної конференції, м. Київ, 18–19 травня 2023 р., с. 126–128. Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського». 2023. URL: https://ktppam.kpi.ua/wp-content/uploads/2023/11/DPTK_blok_tezi_18_05_23_compressed.pdf.
5. Корольова А.В. Мультилінгвальний корпус і його програмне забезпечення для дослідження європеїстики. *Вісник Київського національного лінгвістичного університету. Серія : Філологія,* 2023. *26*(1), 49–62. DOI: 10.32589/2311-0821.1.2023.286184. URL: http://philmessenger.knlu.edu.ua/article/view/286184.
6. Корольова А.В. Мультимедійний підкорпус текстів із європеїстики: критерії для тематичної розмітки. *Науковий вісник Міжнародного гуманітарного університету*. *Серія : Філологія*. 2024. № 68, 87–90. URL: http://www.vestnik-philology.mgu.od.ua/archive/v68/20.pdf; https://doi.org/10.32782/2409- 1154.2024.68.18.
7. Finegan E. LANGUAGE: its structure and use. New York : Harcourt Brace College Publishers. 2004. 613 p.
8. McEnery T. & Wilson A. Corpus Linguistics. Edinburgh : Edinburgh University Press. 2000. 235 p.
9. O'Keeffe A., McCarthy M., & Carter R. From Corpus to Classroom: Language Use and Language Teaching. Cambridge University Press. 2007.

## REFERENCES

1. Zhukovska, V.V. (2015). Corpus linguistics: History of formation and current state. In A.V. Syngaiivska (Ed.). *Current linguistic studies* (textbook) (pp. 168–204). Zhytomyr: Ivan Franko Zhytomyr State University Press.
2. Zhukovska, V.V. (2018). Application of corpus technologies in teaching and learning a foreign language. In Proceedings of the online conference *"Current issues of modern linguistics and methods of teaching language and literature"*, Zhytomyr, February 7–11, 2018 (pp. 39–50). Ivan Franko Zhytomyr State University. Retrieved from: https://nniif.org.ua/File/18zvvzkt.pdf.
3. Zhukovska, V.V. (2020). Linguistic corpus as an innovative informational and research tool of modern linguistics. *Scientific Notes of V.I. Vernadsky Taurida National University. Series: Philology. Social Communications*, 31(70), No. 3, Part 1, 113–119. https://doi.org/10.32838/2663-6069/2020.3-1/20.
4. Zhukovska, V.V. (2023). Corpus technologies and the genre-analytical approach in teaching English for academic purposes. In *Discourse of professional and creative communication: linguocultural, cognitive, translation, and methodological aspects:* Proceedings of the 8th International Scientific and Practical Conference, Kyiv, May 18–19, 2023 (pp. 126–128). National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute". Retrieved from: https://ktppam.kpi.ua/wp-content/uploads/2023/11/DPTK_blok_tezi_18_05_23_compressed.pdf.
5. Korolova, A.V. (2023). Multilingual corpus and its software for European studies research. *Bulletin of Kyiv National Linguistic University. Philology Series*, 26(1), 49–62. https://doi.org/10.32589/2311-0821.1.2023.286184. Retrieved from: http://philmessenger.knlu.edu.ua/article/view/286184.
6. Korolova, A.V. (2024). Multimedia subcorpus of texts on European studies: Criteria for thematic annotation. *Scientific Bulletin of the International Humanities University. Series: Philology*, (68), 87–90. Retrieved from: http://www.vestnik-philology.mgu.od.ua/archive/v68/20.pdf; https://doi.org/10.32782/2409-1154.2024.68.18.
7. Finegan, E. (2004). *Language: Its structure and use*. New York: Harcourt Brace College Publishers.
8. McEnery, T., & Wilson, A. (2000). *Corpus linguistics*. Edinburgh: Edinburgh University Press.
9. O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Дата першого надходження рукопису до видання: 20.06.2025.
Дата прийнятого до друку рукопису після рецензування: 25.07.2025.
Дата публікації: 02.10.2025.