

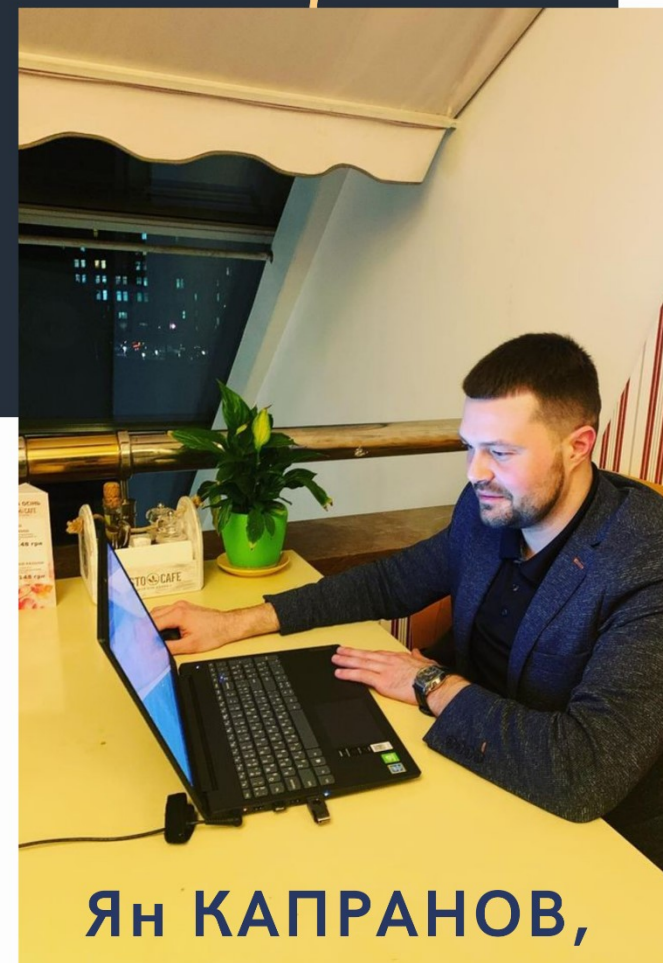


15 лютого 2023 року  
о 13:40 (за київським часом)

# Мультилінгвальні корпусні можливості і ресурси для філологів і перекладачів: європейські практики

## ГОСТЬОВИЙ ВІЗИТ

до Харківського національного університету  
імені В. Н. Каразіна  
(м. Харків, УКРАЇНА)

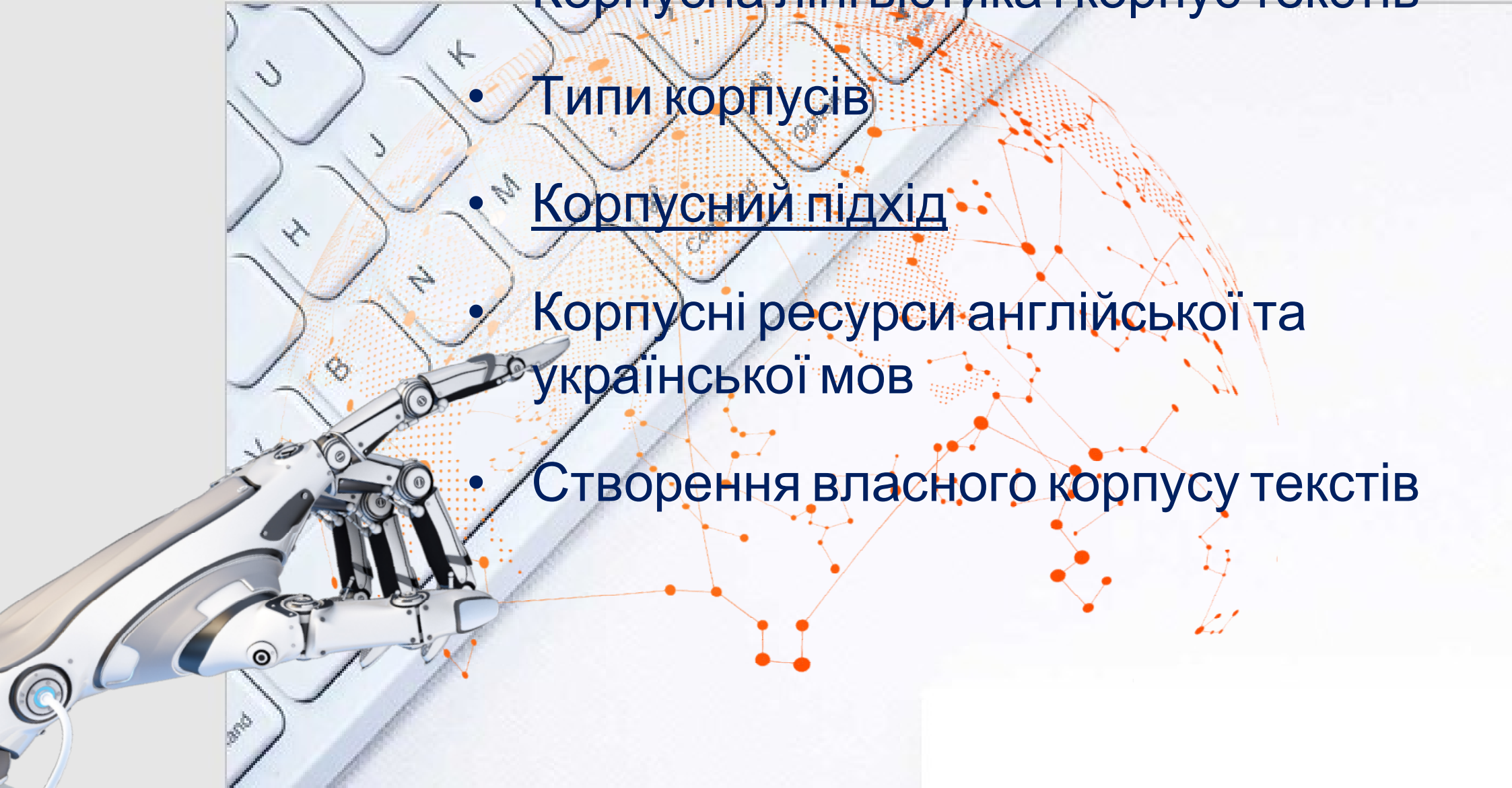


**ЯН КАПРАНОВ,**

доктор філологічних наук, доцент,  
керівник Центру грантової підтримки  
корпусних і перекладацьких досліджень  
Київського національного університету  
лінгвістичного університету (УКРЛІНГ)  
доцент Економіко-гуманітарного факультету  
університету у Варшаві (ПОЛОНІЯ)

# ПЛАН

- Корпусна лінгвістика і корпус текстів
- Типи корпусів
- Корпусний підхід
- Корпусні ресурси англійської та української мов
- Створення власного корпусу текстів



# КЛЮЧОВІ ПОНЯТТЯ

ІНТЕГРАЦІЯ

КОРПУСНА ЛІНГВІСТИКА

КОРПУС ТЕКСТІВ

розділ мовознавства

тексти (письмові й усні)

створення, обробка,  
використання корпусів текстів

певна проблемна галузь

How to do corpus li

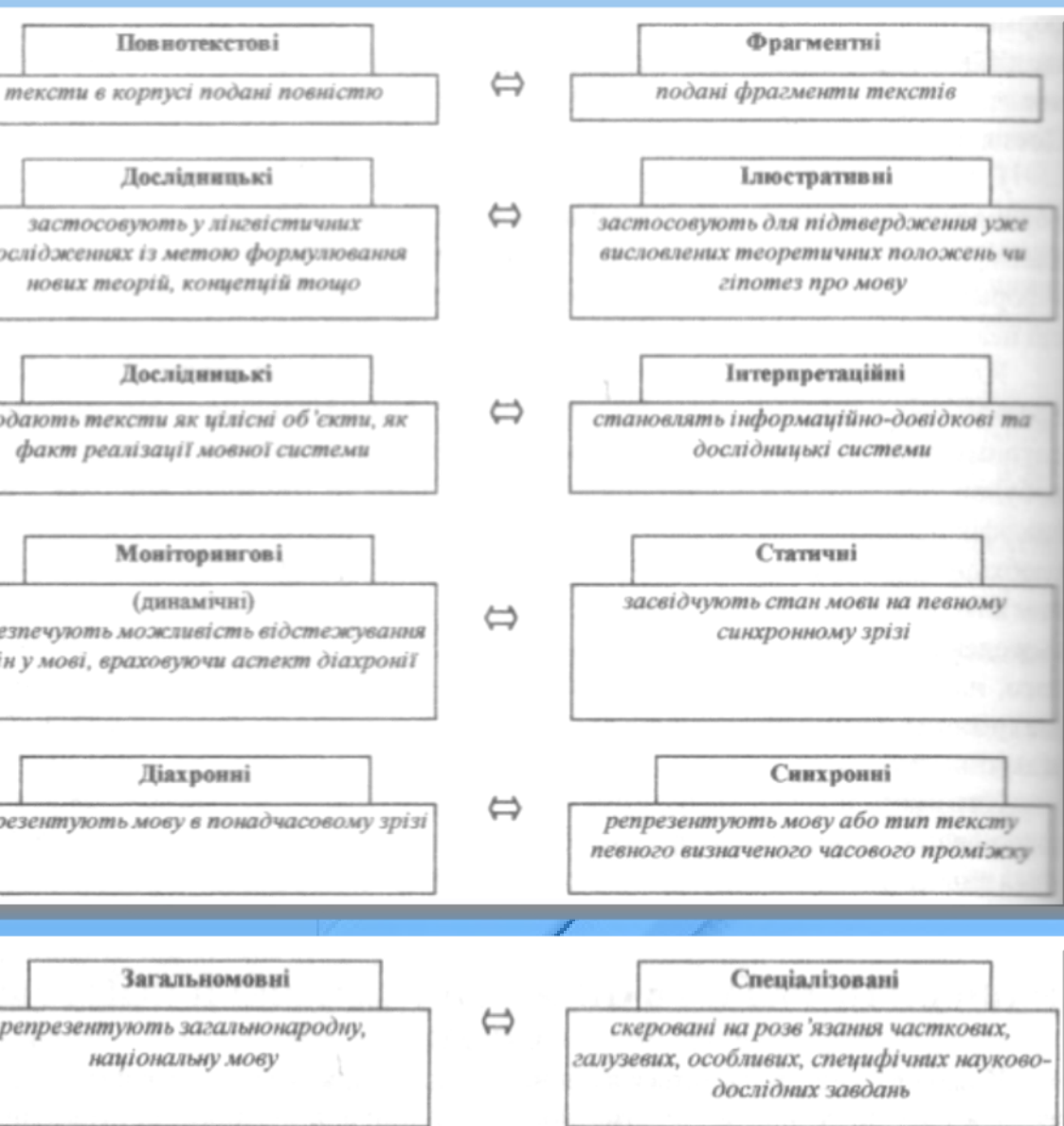


КОМП'ЮТЕР

Спеціалізоване  
програмне  
забезпечення



# ТИПИ КОРПУСІВ



ОЗНАКИ	Типи корпусів
Тип даних	<ul style="list-style-type: none"> <li>•Писемні</li> <li>•Мовні</li> <li>•Змішані</li> </ul>
Мова текстів	<ul style="list-style-type: none"> <li>•Українська</li> <li>•Англійська і т. ін.</li> </ul>
“Паралельність”	<ul style="list-style-type: none"> <li>•Одномовні</li> <li>•Двомовні</li> <li>•Багатомовні</li> </ul>
“Літературність”, специфічність	<ul style="list-style-type: none"> <li>•Літературні</li> <li>•Діалектні</li> <li>•Розмовні</li> <li>•Термінологічні</li> <li>•Змішані</li> </ul>
Жанр	<ul style="list-style-type: none"> <li>•Літературні</li> <li>•Фольклорні</li> <li>•Драматичні</li> <li>•Публіцистичні</li> </ul>
Розмітка	<ul style="list-style-type: none"> <li>•Розмічені</li> <li>•Нерозмічені</li> </ul>
Характер розмітки	<ul style="list-style-type: none"> <li>•Морфологічні</li> <li>•Синтаксичні</li> <li>•Семантичні</li> <li>•Просодичні</li> </ul>

# ТРАНСФЕР КОРПУСНИХ РЕСУРСІВ

у Харківський національний університет імені

В. Н. Каразіна

(м. Харків, УКРАЇНА)

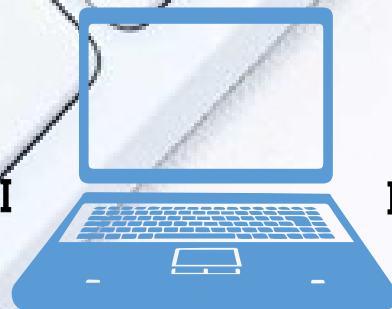
**Data-Driven Learning (DDL)**

=

**Corpus-Aided Discovery Learning**

використання даних корпусів

на заняттях



ТРАДИЦІЙНІ

засоби і

форми

(діалоги,

обговорен

ня,

монологи

тощо)

викладання англійської

та другої іноземної мов

у сучасному європеїзованому світі

ІННОВАЦІЙНІ

І

засоби і

форми

(КОРПУСНИ

Й

ПІДХІД)

ШТУЧНИЙ  
ІНТЕЛЕКТ



СТУДЕНТСЬ  
КА  
АУДИТОРІЯ



# КОМУНІКАТИВНА КОМПЕТЕНТНІСТЬ

## ЗДОБУВАЧ ВИЩОЇ ОСВІТИ – СТУДЕНТ

К о м у н і к а т и в н а  
к о м п е т е н т н і с т ь

### Визначення 1

датність  
ЗВО  
вирішувати  
актуальні  
завдання  
спілкування  
засобами  
іноземних  
мов  
(англійсько  
ю, німецькою

### Визначення 2

вміння  
користуват  
ися фактами  
мови та мови  
для  
реалізації  
цілей  
спілкування,  
тощо



# Призначення корпусу

частота граматичних категорій,  
слів, словосполучень, тощо.

РЕПРЕЗЕНТАТИВ  
НІСТЬ

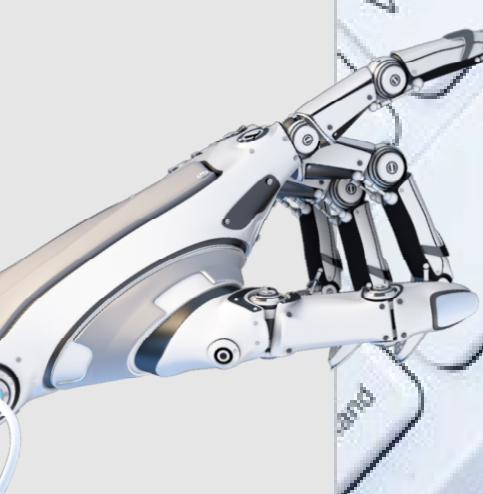
комбінаторика  
мовних одиниць

Призначення  
корпусу

зміна  
частот

зміна контекстів  
у різні періоди часу

статус мовних одиниць  
у різних авторів



: What would be the value of the goods and services  
gers in 2010 to compute the value of goods and services  
55 level. Put differently, the output of goods and services  
n particular, GDP omits the value of goods and services  
of the market value of all the goods and services  
is the market value of ail final goods and services  
ments. Net exports equal the value of goods and services  
abroad (exports) minus the value of goods and services  
GDP deflator reflects the prices of all goods and services  
fers from the CP! because it includes goods and services  
e role of productivity-the amount of goods and services  
a word, productivity, the quantity of goods and services  
ause it measures the total quantity of goods and services

# Конкорданс

## ПОШУКОВА СИСТЕМА

### ВИЗНАЧЕННЯ

#### Визначення 1

спеціалізована  
лінгвістична  
прикладна  
програма

#### Визначення 2

автоматична  
вибірка  
заданих  
мовних  
одиниць

електронних  
текстів

N	Concordance
1	Petroleum Authority (NPA) states that the debt is a
2	by the Committee and touch on two points. A
3	memory, the clause 8 that they have called is a
4	it I would be able to speak to the facts. But I a
5	given the interest free loan of GH¢55 million. He a
6	know, for instance, that the Bui Authority, which a
7	, such that they are not happy and had a
8	price of US\$3,000 per metric tonne. We are a
9	on Food, Agriculture and Cocoa Affairs a
10	â€œcreate, loot and shareâ€ agenda? We have a



# Приклади корпусних ресурсів англійської та української мов

**BRITISH NATIONAL CORPUS**

**About**

- What is the BNC?
- Creating the BNC
- BNC Products
- Copyright
- Contact Us
- Contents A-Z

**About the BNC**

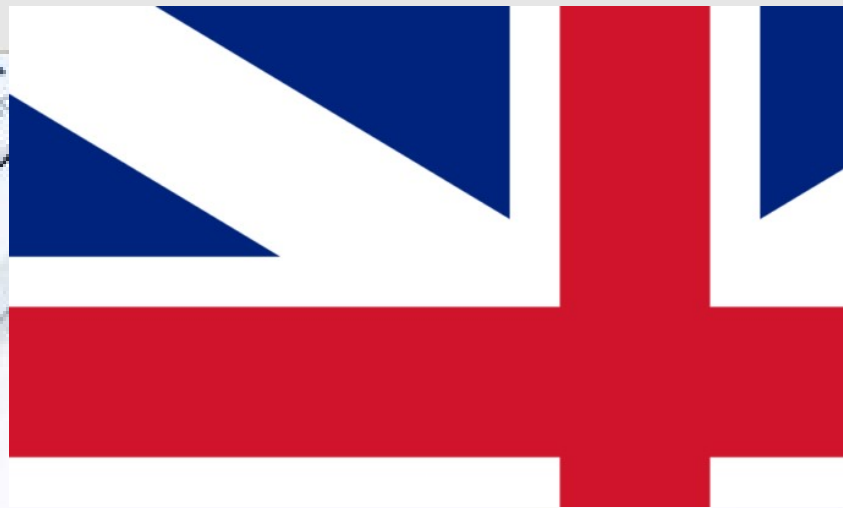
The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. [\[more\]](#)

**Simple Search from the British Library**

Type a word or phrase in the search box and press the Go button to see up to 50 random hits from the corpus.

Look up:

You can search for a single word or a phrase, restrict searches by



Генеральний регіонально анотований корпус української мови (ГР)  
М. Шведова, Р. фон Вальденфельс, С. Яригін, А. Рисін, В. Старко,  
Ніколаєнко та ін. – Київ, Львів, Єна, 2017-2022. – uacorporus.org

**DASHBOARD** Grac v.12

**GRAC V.12**

- Word Sketch: Collocations and word combinations
- Thesaurus: Synonyms and similar words
- Parallel Concordance: Translation search
- N-grams: Multword expressions (MWEs)
- Trends: Diachronic analysis, neologisms

**Word Sketch Difference**: Compare collocations of two words

**Concordance**: Examples of use in context

**Wordlist**: Frequency list

**Keywords**: Terminology extraction


**One-Click Dictionary**: Automatic dictionary drafting








**RECENTLY USED CORPORA**

Corpus	Language	Count
Grac v.12	Ukrainian	640
Grac v.11	Ukrainian	598
Grac v.10	Ukrainian	598
Grac v.12	Ukrainian	640
Grac v.11	Ukrainian	598
Grac v.12	Ukrainian	640
Grac v.11	Ukrainian	598
Grac v.10	Ukrainian	598
Grac v.11	Ukrainian	598
Grac v.12	Ukrainian	640
Grac v.12	Ukrainian	640
Grac v.12	Ukrainian	640

Master the interface in 2 days!  
March & April 2020

## Приклади корпусних ресурсів англійської мов

The links below are for the online interface. But you can also  download the corpora for use

Corpus (online access)	Download	# words	Di
<a href="#">iWeb: The Intelligent Web-based Corpus</a>		14 billion	6 co
<a href="#">News on the Web (NOW)</a>		8.7 billion+	20 cc
<a href="#">Global Web-Based English (GloWbE)</a>		1.9 billion	20 cc
<a href="#">Wikipedia Corpus</a>		1.9 billion	(Va
<a href="#">Corpus of Contemporary American English (COCA)</a>		560 million	Am
<a href="#">Corpus of Historical American English (COHA)</a>		400 million	Am
<a href="#">The TMC</a>		225 million	6

# Корпусний ресурс

АНГЛІЙСЬКА МОВА

англ

КОМУНІКАЦІЯ  
на прикладі  
реального  
спілкування

***N-grams*** are sequences of  $n$  words, where  $n$  falls in the range 1-8, and *word* means a token of any lexical item assigned a PoS tag by the CLAWS parser ([details](#)). For example, the most frequent 1-gram in the BNC is *the*, and *the end of the* tops the list of 4-grams.

***Phrase-frames*** are **sets of variants** of an  $n$ -gram identical except for one word, represented here by the word symbol \*. The most frequent (and most productive, *i.e.* having the greatest number of variants) 4-frame is *the \**, with 5652 variants such as *the end of the*, *the rest of the*, *the top of the*, *the nature of the* etc.

***PoS-grams*** are patterns of **Part of Speech** tags assigned to word forms without reference to the specific entities. When ordered by **types**, the most frequent "3-PoS-gram" is ART ADJ NOUN, e.g. *the other hand*. On the other hand, when ordered by **tokens**, the 3-PoS-gram PREP ART NOUN as in *at the end* are more frequent.

***Char-grams*** are sequences of  $n$  letters. Their distribution can be studied by position (initial, medial, final) as well as by frequency in tokens or types. Unsurprisingly, *the* is the most frequent 3-char-gram by tokens (8,210 tokens, 1007 types), but *ing* has the most distinct types (2,991,683 tokens, 9416 types).

## BRITISH NATIONAL CORPUS

### About

- [What is the BNC?](#)
- [Creating the BNC](#)
- [BNC Products](#)
- [Copyright](#)
- [Contact Us](#)
- [Contents A-Z](#)

### Using the BNC

### About the BNC

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. [\[more\]](#)

### Simple Search from the British Library

Type a word or phrase in the search box and press the Go button to see up to 50 random hits from the corpus.

Look up:

You can search for a single word or a phrase, restrict searches by

BRITISH  
NATIONAL  
CORPUS

**The British National Corpus:** The platform gives access to five million words from the BNC representing informal conversations between British English speakers from the 1990s.

BNC  
BRITISH  
NATIONAL  
CORPUS

**The British National Corpus 2014:** The platform gives access to five million words from the BNC 2014 representing informal conversation between British English speakers from 2000s.

# Корпусний ресурс англійської мови

BRITISH NATIONAL CORPUS

## About the BNC

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. [\[more\]](#)

## Simple Search from the British Library

Type a word or phrase in the search box and press the Go button to see up to 50 random hits from the corpus.

Look up:

You can search for a single word or a phrase, restrict searches by

**Exercise 1. Provide the Frequency for the Words, their Synonyms, and Antonyms from British National Corpus.**

Words	Synonyms – Same Meaning	Antonyms – Opposites
adhere	comply, observe	condemn, disjoin
abolish	abrogate, annul	setup, establish
alien	foreigner, outsider	native, resident
awkward	rude, blundering	adroit, clever
busy	active, engaged	idle, lazy
comprise	include, contain	reject, lack
calm	harmonious, unruffled	stormy, turbulent

# Корпусний ресурс української мови

Генеральний регіонально анотований корпус української мови (ГРАК) / М. Шведова, Р. фон Вальденфельс, Р. Фігін, А. Рисін, В. Старко, Т. Ніколаєнко та ін. – Київ, Львів, Єна, 2017-2022. – uacorpus.org.



Завдання 2. Перевірте частотність уживання поданих українських відповідників за корпусними ресурсами української мови

The screenshot shows the Sketch Engine interface with the search term 'ride' entered. The results section shows the lemma 'ride' with a frequency of 15,703,895,409. The interface includes various tool panels like Word Sketch, Thesaurus, and Wordlist.

Word Combinations	Ukrainian Translation	Frequency
get out of hand	вийти з-під контролю	
	інші варіанти	
so far so good	все йде нормально	
	все чудово	
	поки все в порядку	
	інші варіанти	
pull yourself together	взяти себе в руки	
	інші варіанти	

ВОДЧИК

Текст

Документы

Сайты

ОПРЕДЕЛИТЬ ЯЗЫК

АНГЛИЙСКИЙ

РУССКИЙ

НЕМЕЦКИЙ



УКРАИНСКИЙ

strong



СИЛЬНИЙ

strôNG



6 / 5 000



syl'nyy



strong – определения

Имя прилагательное

- 1 having the power to move heavy weights or perform other physically demanding tasks.  
“she cut through the water with her strong arms”

Синонимы:

strong: вариан

Имя прилагате

вигривалий

дужий

міцний



# Створення власного

## корпусу текстів

AntConc

File Edit Settings Help

**Target Corpus**  
Name: AmE06\_Learned  
Files: 80  
Tokens: 161469

**Reference Corpus**  
Name: AmE06  
Files: 500  
Tokens: 1017879

Progress 100%

KWIC Plot File Cluster N-Gram Collocate Word Keyword

Total Hits: 42 Page Size 100 hits 1 to 42 of 42 hits

	File	Left Context	Hit	Right Context
1	AmE06_J32.txt	ing word is a verb, but not when the -ing	word	is a noun or adjective. So the grammar of
2	AmE06_J47.txt	object, action or quality. Learning the category boundaries for each	word	is a specific "problem of induction." Children are placed
3	AmE06_J32.txt	a-hunting dog. The a- is possible when the -ing	word	is a verb, but not when the -ing word
4	AmE06_J34.txt	which coincides with word-initial position, and that the entire	word	is dominated by a single syllable. The gesture-calculations
5	AmE06_J59.txt	as Marjorie Perloff puts it, or perhaps that they put	word	and image in a mutually interpretive context, as the
6	AmE06_J59.txt	Stein's verbal portraits of these painters "attempt to fuse	word	and image," as Marjorie Perloff puts it, or perhaps
7	AmE06_J60.txt	an ancient purity and directness, a pre-Babelic unity of	word	and thing. In so doing, though, photography was just
8	AmE06_J47.txt	the word, or if the child understands and says the	word.	The second component is a list of 63 communicative, social,
9	AmE06_J47.txt	to indicate if they have heard their child say the	word.	The second major component utilizes an innovative sentence pair
10	AmE06_J47.txt	Consider what would appear to be the simplest condition of	word	learning. A fluent speaker of the language, such as
11	AmE06_J47.txt	MCDI and laboratory measures of comprehension, production, and	word	learning. The second is the Twins' Early Development Study (
12	AmE06_J47.txt	Each type of reference is a plausible use of the	word.	And even if we were correct in believing that "

Search Query  Words  Case  Regex Results Set All hits Context Size 10 token(s)

word Start  Adv Search

Sort Options Sort to right Sort 1 1R Sort 2 2R Sort 3 3R Order by freq

Time taken (creating KWIC display): 0.0043 ms



The background is a dark blue space scene featuring a constellation of stars connected by thin white lines. Two glowing blue wireframe cubes are positioned in the upper right, one slightly behind and to the right of the other, creating a sense of depth. The text "ДЯКУЮ ЗА УВАГУ!" is centered in the lower half of the image in a white, sans-serif font.

ДЯКУЮ ЗА УВАГУ!