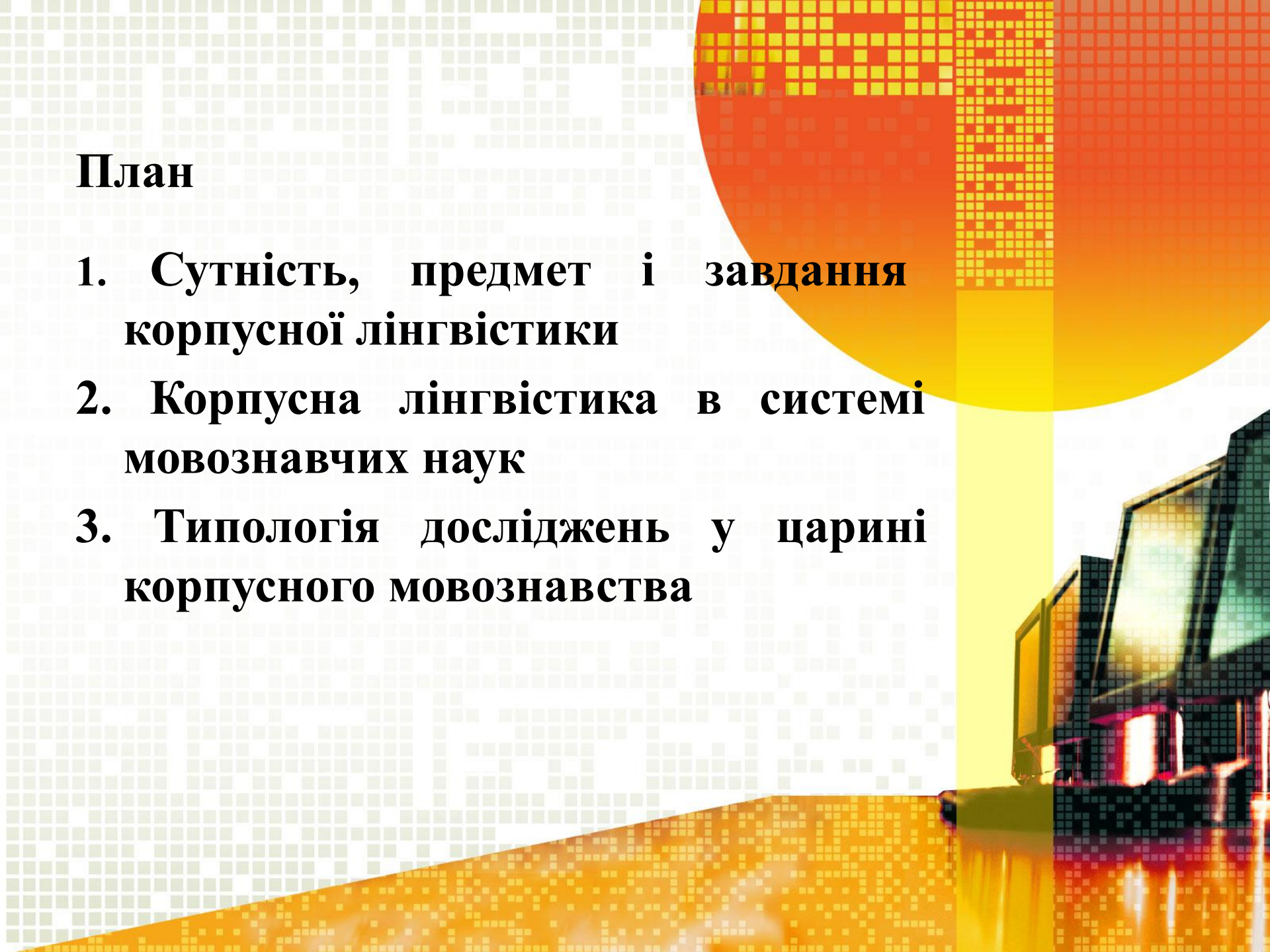
Лекція 1

Корпусна лінгвістика як галузь прикладного мовознавства

План

1. **Сутність, предмет і завдання корпусної лінгвістики**
2. **Корпусна лінгвістика в системі мовознавчих наук**
3. **Типологія досліджень у царині корпусного мовознавства**



1. Сутність, предмет і завдання корпусної лінгвістики

Корпусна лінгвістика - це нова лінгвістична галузь, що розпочала своє активне становлення у 60-х роках XX століття у зв’язку із інтенсивним розвитком комп’ ютерних технологій.

Термін «корпусна лінгвістика» - XX століття з публікацією у 1983 році збірника наукових праць «Corpus Linguisitcs: Recent Developments in the Use of Computer Corpora in English Language Research».

Корпусна лінгвістика займається

- **визначенням загальних принципів побудови, обробки та експлуатації даних лінгвістичних корпусів (корпусів текстів) із використанням сучасних комп’ ютерних технологій;**

- **розробленням методики збору реальних мовних явищ – писемних та усних текстів, а також способів їх збереження та аналізу.**

Корпус текстів **- це значний за обсягом, представлений в електронному вигляді, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, створений для вирішення конкретних лінгвістичних завдань [Захаров, 2005: 3].**

BYU-BNC: BRITISH NATIONAL CORPUS

100 MILLION WORDS, 1980s-1993

TORY

history | lists | profile | log

DISPLAY

LIST CHART KWIC COMPARE

SEARCH STRING

WORD(S)

COLLOCATES

POS LIST

RANDOM SEARCH RESET

SECTIONS SHOW

1 IGNORE ----- SPOKEN FICTION MAGAZINE NEWSPAPER NON-ACAD

2 IGNORE ----- SPOKEN FICTION MAGAZINE NEWSPAPER NON-ACAD

SORTING AND LIMITS

SORTING FREQUENCY

MINIMUM FREQUENCY 5

CLICK TO SEE OPTIONS

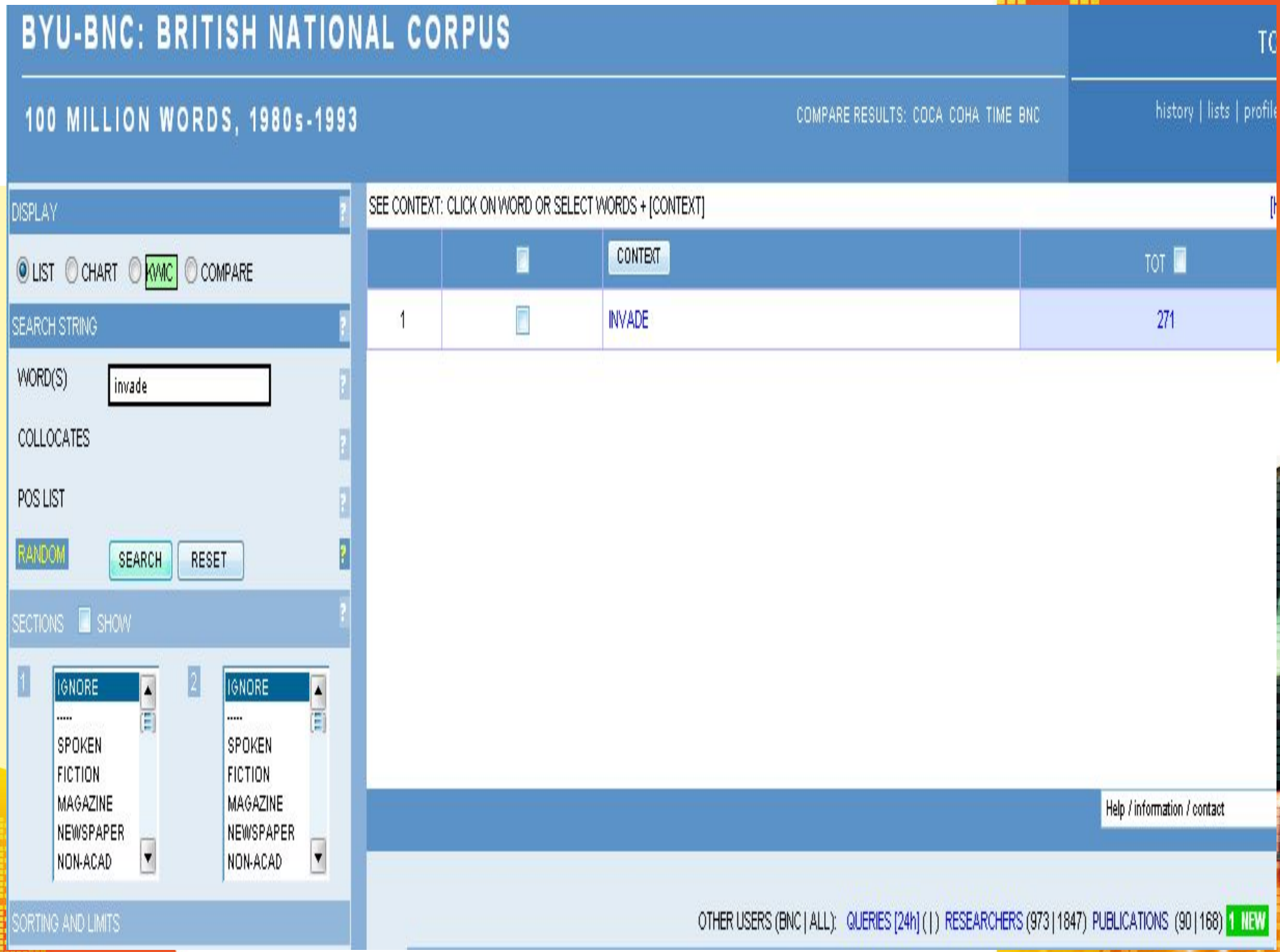
SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC

OVERVIEW: VICTORIA ZHUKOVSKAYA

Help / information / contact

OTHER USERS (BNC | ALL): QUERIES [24h] (|) RESEARCHERS (973 | 1847) PUBLICATIONS (90 | 168) 1 NEW

Personal information	
Category	Faculty (Professor/Lecturer) Update
Institution	zhitomir state university Update
Country	UKRAINE Update
Short profile (edit)	teaching English as a second foreign language
Status	
Your publications	0
Access level	3 / 5
Usage limits	CLICK TO SEE
Activity	


BYU-BNC: BRITISH NATIONAL CORPUS
100 MILLION WORDS, 1980s-1993
COMPARE RESULTS: COCA COHA TIME BNC
history | lists | profile
DISPLAY
LIST CHART KWIC COMPARE
SEARCH STRING
WORD(S)
invade
COLLOCATES
POS LIST
RANDOM
SEARCH
RESET
SECTIONS SHOW
1
IGNORE
.....
SPOKEN
FICTION
MAGAZINE
NEWSPAPER
NON-ACAD
2
IGNORE
.....
SPOKEN
FICTION
MAGAZINE
NEWSPAPER
NON-ACAD
SORTING AND LIMITS
SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT]
CONTEXT
TOT
1
INVADE
271
Help / information / contact
OTHER USERS (BNC | ALL): QUERIES [24h] (|) RESEARCHERS (973 | 1847) PUBLICATIONS (90 | 168) 1 NEW

BYU-BNC: BRITISH NATIONAL CORPUS

100 MILLION WORDS, 1980s-1993

COMPARE RESULTS: COCA COHA TIME BNC

TORY

history | lists | profile | lo

Help / information / contact

KEYWORD IN CONTEXT DISPLAY

SECTION: NO LIMITS

PAGE: << < 1 / 3 > >>
SAMPLE: 100 200

CLICK FOR MORE CONTEXT [?] SAVE LIST CHOOSE LIST ..... CREATE NEW LIST [?]

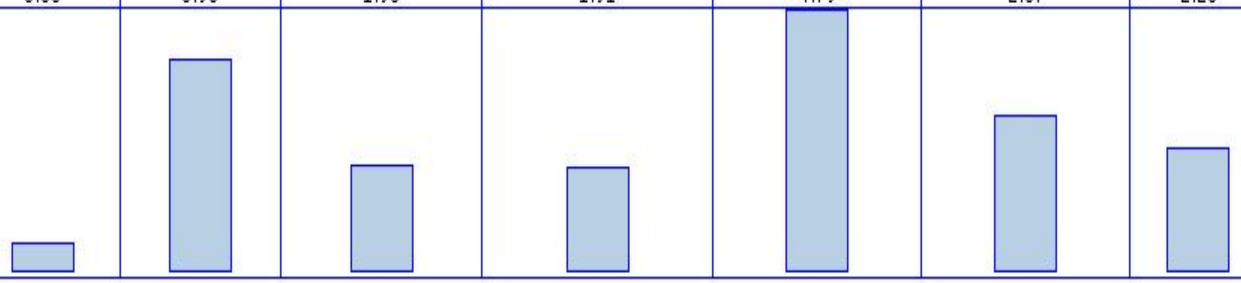
1	JSK	S_lect_polit_law_edu	A	B	C	Americans intervene there, nor could they cross the demilitarized zone in force and erm **invade** North er Vietnam as they had done with North Korea previously. Now these limits
2	HVJ	S_pub_debate	A	B	C	's due for discussion on policy H Two, that issue does shall I suggest **invade** policy H One, because an appropriate form of words has to be agreed in
3	HNP	W_fict_poetry	A	B	C	seat of his favourite cafe. He was convalescing, weightless, letting the steam **invade** his thought, forgetting the mountains, the words on vellum, his body anchored
4	CH5	W_newsp_tabloid	A	B	C	a kong is four of a kind, and a chow is a run. **IIIVADE**, writes: BETWEEN the outbreak of war and the Nazi invasion of the Low
5	CH5	W_newsp_tabloid	A	B	C	. If that happens the aliens will be able to disguise themselves as humans and **invade** Earth. Nothing less than the future of Mankind is at stake! Armed with
6	CH6	W_newsp_tabloid	A	B	C	Brothers Phil and Fergus Gribbon, of Co Donegal, have hired a trawler to **invade** the uninhabited isle of Rockall, 300 miles west of the Hebrides. They plan
7	A4C	W_newsp_brdsht_nat_misc	A	B	C	them -- sometimes successfully -- in conversations about the need for change. Pin-headed aliens **invade** Russia -- official From RUPERT CORNWELL in Moscow NIGHT LIFE in Voronezh, a dour
8	AAF	W_newsp_brdsht_nat_misc	A	B	C	inspiration in the carving, but that a new era of peace and honesty should **invade** my life whilst I worked... " A later entry reveals an undesirable recognition of
9	A9N	W_newsp_brdsht_nat_report	A	B	C	be banned, or the precedent would allow a flood of other charitable panhandlers to **invade** their territory, and solicit a little of the money which should by rights be the
10	AAU	W_newsp_brdsht_nat_report	A	B	C	to ensure that the 1,400 journalists and hundreds of European officials who are expected to **invade** the capital for the June summit depart with a good impression of Irish hospitality.
11	AL2	W_newsp_brdsht_nat_commerce	A	B	C	gold to twitch up like a kneecap tapped with a hammer. Now Saddam can **invade** , the Soviet Union can fall apart and Japanese asset values can collapse without gold
12	CBC	W_newsp_other_social	A	B	C	check-out staff are on full Euro alert. The British are coming, preparing to **invade** the Continent armed with open wallets and ready-to-fill shopping bags. When the European trade
13	CEK	W_newsp_other_social	A	B	C	In some cases, I'm told, pieces of the womb break free and **invade** other parts of the body, including the lungs, so it's best not
14	K5J	W_newsp_other_social	A	B	C	habitually cheat, or the spectators who take the law into their own hands and **invade** football pitches. At the weekend, the world indoor athletics championships took place in
15	CEN	W_newsp_other_report	A	B	C	the Ulster Defence Regiment, had been found guilty of manslaughter. French noisy frogs **invade** LOUD French frogs are disturbing the sleep of English home owners. A Wildlife Trust
16	CEN	W_newsp_other_report	A	B	C	the site of a massive rave party. Thousands of youngsters are tonight expected to **invade** the grounds of the £7,000-a-year Cobham Hall School near Gravesend, Kent. Only headmistress
17	K5M	W_newsp_other_report	A	B	C	," she said. Paul received information from other hackers which he used to **invade** the confidential files of international institutions. Only sheer exhaustion would force Paul to stop
18	K5M	W_newsp_other_report	A	B	C	reaction in the North, which suspects that the US and Seoul are planning to **invade** under the guise of war games. Indeed, there must be serious doubts about
19	CEP	W_newsp_other_sports	A	B	C	by beating Leeds to progress to the Champions League, begin their quest to successfully **invade** Europe against Marseille at Ibrox tonight. But their captain, Richard Gough, aware
20	FP8	W_ac_polit_law_edu	A	B	C	are certain rights which citizens have and which the legislature or the executive must not **invade** or remove. They may feel that certain laws should not be made at all...
21	HXF	W_ac_polit_law_edu	A	B	C	dilemma that they have to solve or explain: Why did the Romans decide to **invade** the island of Britain? It's much harder conquering an island. (They

CLICK ON BARS FOR CONTEXT

CLICK ON COLUMN HEADINGS FOR FREQUENCY IN SUB-SECTION

SECTION	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC
FREQ	5	62	14	20	79	44	47
PER MIL	0.50	3.90	1.93	1.91	4.79	2.87	2.26

SEE ALL SUB-SECTIONS AT ONCE



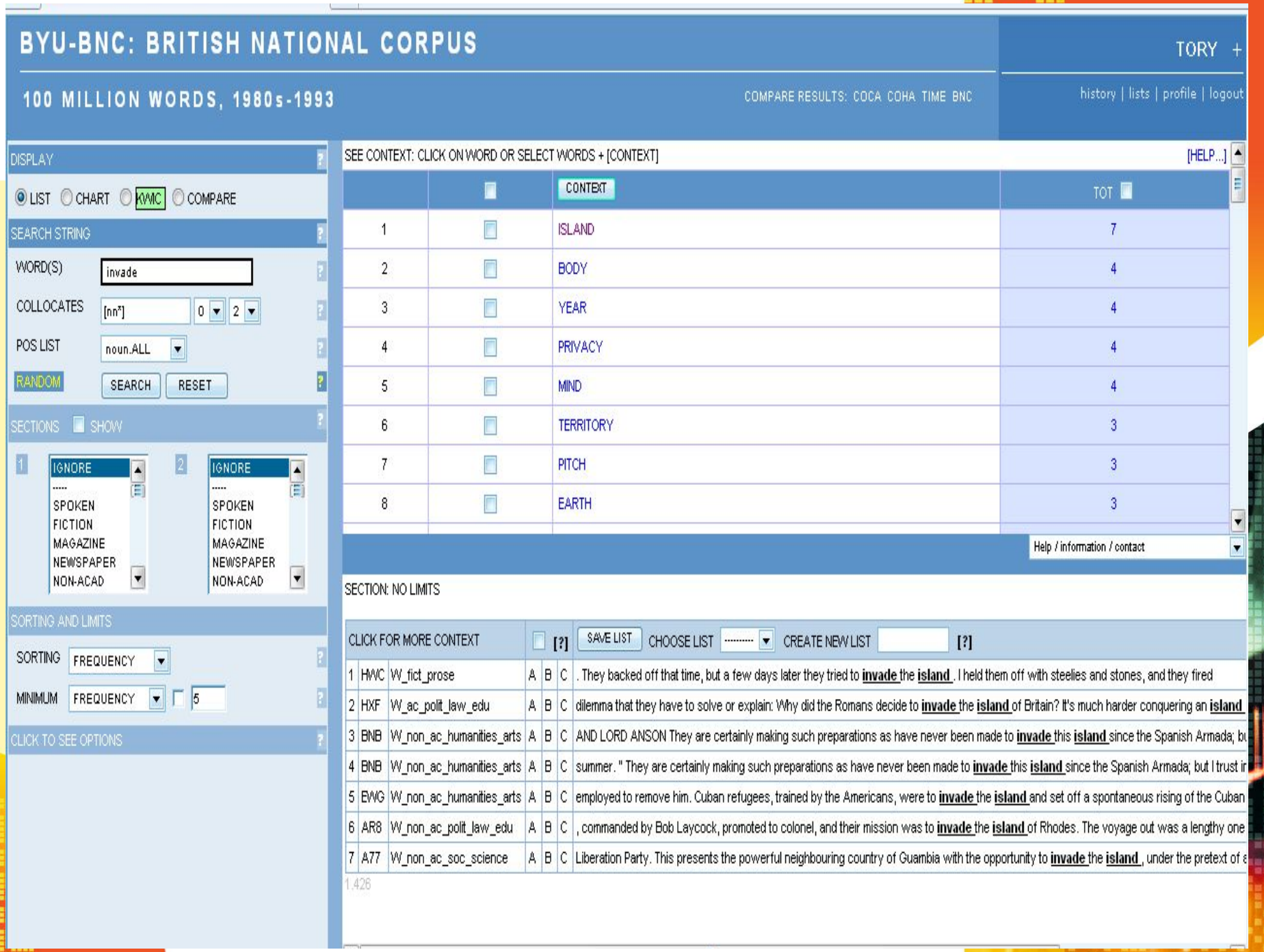
Help / information / contact

SECTION: NO LIMITS

CLICK FOR MORE CONTEXT [?] SAVE LIST CHOOSE LIST --------- CREATE NEW LIST [?]

1	HWC	W_fict_prose	A	B	C	. They backed off that time, but a few days later they tried to **invade** the **island** . I held them off with steelies and stones, and they fired
2	HXF	W_ac_polit_law_edu	A	B	C	dilemma that they have to solve or explain: Why did the Romans decide to **invade** the **island** of Britain? It's much harder conquering an **island** . (They
3	BNB	W_non_ac_humanities_arts	A	B	C	AND LORD ANSON They are certainly making such preparations as have never been made to **invade** this **island** since the Spanish Armada; but I trust in God and Lord Anson
4	BNB	W_non_ac_humanities_arts	A	B	C	summer. " They are certainly making such preparations as have never been made to **invade** this **island** since the Spanish Armada; but I trust in God and Lord Anson
5	EWG	W_non_ac_humanities_arts	A	B	C	employed to remove him. Cuban refugees, trained by the Americans, were to **invade** the **island** and set off a spontaneous rising of the Cuban people against Castro.
6	AR8	W_non_ac_polit_law_edu	A	B	C	, commanded by Bob Laycock, promoted to colonel, and their mission was to **invade** the **island** of Rhodes. The voyage out was a lengthy one via the Cape
7	A77	W_non_ac_soc_science	A	B	C	Liberation Party. This presents the powerful neighbouring country of Guambia with the opportunity to **invade** the **island** , under the pretext of a request for assistance from the royal family

1.426


BYU-BNC: BRITISH NATIONAL CORPUS
100 MILLION WORDS, 1980s-1993
COMPARE RESULTS: COCA COHA TIME BNC
TORY +
history | lists | profile | logout
DISPLAY
LIST CHART KWIC COMPARE
SEARCH STRING
WORD(S) invade
COLLOCATES [nn*] 0 2
POS LIST noun.ALL
RANDOM SEARCH RESET
SECTIONS SHOW
1 IGNORE ----- SPOKEN FICTION MAGAZINE NEWSPAPER NON-ACAD
2 IGNORE ----- SPOKEN FICTION MAGAZINE NEWSPAPER NON-ACAD
SORTING AND LIMITS
SORTING FREQUENCY
MINIMUM FREQUENCY 5
CLICK TO SEE OPTIONS
SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT]
[HELP...]
CONTEXT
TOT
1 ISLAND 7
2 BODY 4
3 YEAR 4
4 PRIVACY 4
5 MIND 4
6 TERRITORY 3
7 PITCH 3
8 EARTH 3
Help / information / contact
SECTION: NO LIMITS
CLICK FOR MORE CONTEXT [?] SAVE LIST CHOOSE LIST ---------- CREATE NEW LIST [?]
1 HWC W_fict_prose A B C . They backed off that time, but a few days later they tried to invade the island . I held them off with steelies and stones, and they fired
2 HXF W_ac_polit_law_edu A B C dilemma that they have to solve or explain: Why did the Romans decide to invade the island of Britain? It's much harder conquering an island
3 BNB W_non_ac_humanities_arts A B C AND LORD ANSON They are certainly making such preparations as have never been made to invade this island since the Spanish Armada; bu
4 BNB W_non_ac_humanities_arts A B C summer. " They are certainly making such preparations as have never been made to invade this island since the Spanish Armada; but I trust ir
5 EWG W_non_ac_humanities_arts A B C employed to remove him. Cuban refugees, trained by the Americans, were to invade the island and set off a spontaneous rising of the Cuban
6 AR8 W_non_ac_polit_law_edu A B C , commanded by Bob Laycock, promoted to colonel, and their mission was to invade the island of Rhodes. The voyage out was a lengthy one
7 A77 W_non_ac_soc_science A B C Liberation Party. This presents the powerful neighbouring country of Guambia with the opportunity to invade the island , under the pretext of
1,426

BYU-BNC: BRITISH NATIONAL CORPUS

TORY +

100 MILLION WORDS, 1980s-1993

COMPARE RESULTS: COCA COHA TIME BNC

history | lists | profile | logout

DISPLAY

LIST CHART KWIC COMPARE

SEARCH STRING

WORD(S) invade attack

COLLOCATES

POS LIST

RANDOM SEARCH RESET

SECTIONS SHOW

1 IGNORE ----- SPOKEN FICTION MAGAZINE NEWSPAPER NON-ACAD

2 IGNORE ----- SPOKEN FICTION MAGAZINE NEWSPAPER NON-ACAD

SORTING AND LIMITS

SORT BY FREQUENCY WORD 1 : 2

MINIMUM FREQUENCY 5 5

CLICK TO SEE OPTIONS

SEE CONTEXT: CLICK ON NUMBERS (WORD 1 OR 2) [HELP...]

WORD 1 (W1): **IINVADE** (0.03)

	WORD	W1	W2	W1/W2	SCORE
1	THE	72	1713	0.0	1.4
2	.	38	1561	0.0	0.8
3	,	24	982	0.0	0.8
4	AND	17	429	0.0	1.4
5	ENGLAND	13	6	2.2	74.2
6	HER	11	62	0.2	6.1
7	THIS	8	60	0.1	4.6
8	ISLAND	7	0	14.0	479.2
9	THEIR	7	61	0.1	3.9
10	HIS	7	88	0.1	2.7

WORD 2 (W2): **ATTACK** (34.23)

	WORD	W2	W1	W2/W1	SCORE
1	ON	1923	1	1,923.0	56.2
2	THE	1713	72	23.8	0.7
3	.	1561	38	41.1	1.2
4	,	982	24	40.9	1.2
5	A	481	3	160.3	4.7
6	AND	429	17	25.2	0.7
7	IN	412	5	82.4	2.4
8	BY	323	2	161.5	4.7
9	"	294	6	49.0	1.4
10	OF	272	0	544.0	15.9

Help / information / contact

SECTION: NO LIMITS

CLICK FOR MORE CONTEXT [?] SAVE LIST CHOOSE LIST ---------- CREATE NEW LIST [?]

1	HWC	W_fict_prose	A	B	C	. They backed off that time, but a few days later they tried to **invade** the **island** . I held them off with steelies and stones, and they fired
2	HXF	W_ac_polit_law_edu	A	B	C	dilemma that they have to solve or explain: Why did the Romans decide to **invade** the **island** of Britain? It's much harder conquering an **island**
3	BNB	W_non_ac_humanities_arts	A	B	C	AND LORD ANSON They are certainly making such preparations as have never been made to **invade** this **island** since the Spanish Armada; bu
4	BNB	W_non_ac_humanities_arts	A	B	C	summer. " They are certainly making such preparations as have never been made to **invade** this **island** since the Spanish Armada; but I trust ir
5	EWG	W_non_ac_humanities_arts	A	B	C	employed to remove him. Cuban refugees, trained by the Americans, were to **invade** the **island** and set off a spontaneous rising of the Cuban
6	AR8	W_non_ac_polit_law_edu	A	B	C	, commanded by Bob Laycock, promoted to colonel, and their mission was to **invade** the **island** of Rhodes. The voyage out was a lengthy one
7	A77	W_non_ac_soc_science	A	B	C	Liberation Party. This presents the powerful neighbouring country of Guambia with the opportunity to **invade** the **island** , under the pretext of a

1,426

BYU-BNC: BRITISH NATIONAL CORPUS

TORY

100 MILLION WORDS, 1980s-1993

COMPARE RESULTS: COCA COHA TIME BNC

history | lists | profile | lo

Help / information / contact

DIS

SEA STR

W (S)

C

PC LIS

RA

SEC SH

1

DIS SO

DI

SC

CL TO OP

KEYWORD IN CONTEXT DISPLAY

1	ABD	W_pop_lore	A	B	C	foreign companies are likely to bid for the rare chance to	invade	a new broadcasting market . Until now British TV licences were
2	G1H	W_ac_soc_science	A	B	C	the result and one side 's league position . Whereas to	invade	a pitch when a goal is disallowed may be felt magically as
3	H0A	W_biography	A	B	C	had served under Gordon in Equatoria , had set out to	invade	Abyssinia from Tajura , and been exterminated by the Danakil
4	JXW	W_fict_prose	A	B	C	rasped , cupping her face , drawing her inexorably back to	invade	again the full softness of her mouth . Her senses felt drugged
5	G00	W_commerce	A	B	C	instant transformations from one shape to another will start to	invade	all sorts of artwork before very long ! It 's certainly fun
6	ABK	W_pop_lore	A	B	C	and First Fidelity Into the breach ARE European banks about to	invade	America -- again ? On March 19th Banco de Santander , Spain
7	HA5	W_fict_prose	A	B	C	need , his mounting passion , his overriding desire to	invade	and conquer her . As his firm , predatory fingers stroked her
8	CM4	W_fict_prose	A	B	C	by which the malevolent lunacy of powers in the warp could	invade	and ravage worlds ; could corrupt the human race into polluted
9	B77	W_non_ac_nat_science	A	B	C	has been aptly described as an act of hijack . Viruses	invade	animal , plant and bacterial cells , and commandeer the complex
10	FF0	W_non_ac_nat_science	A	B	C	dirty stories , and what you learn there will mysteriously	invade	any other field in which timing is important , unless some mental
11	ALY	W_non_ac_humanities_arts	A	B	C	but Winston Churchill did not trust Hitler and defied him to	invade	Britain . Appreciating that Britain had become even more
12	F9H	W_misc	A	B	C	century the Spanish Armada , before its abortive attempt to	invade	Britain . was ordered to destroy . if nothing else . the
13	CM6	W_ac_humanities_arts	A	B	C	obvious that Austria-Hungary would not allow the Russians to	invade	Bulgaria . a year before she had disliked the possibility of a
14	G1R	W_ac_humanities_arts	A	B	C	" power , he did secure an American guarantee not to	invade	Cuba . which has diminished the spectre of how the Soviet Union
15	EWG	W_non_ac_humanities_arts	A	B	C	missiles in return for a public pledge that the USA would not	invade	Cuba . This conciliatory letter was followed by a sterner
16	HR7	W_fict_prose	A	B	C	Tall " reading an illustrated journal called Voodoo monsters	invade	earth . " You saw the operational machines , " said Harvey
17	H7F	W_fict_prose	A	B	C	lot of weird aliens called the Sproati and they decide to	invade	Earth " " I think this has been done before , "
18	G16	W_fict_prose	A	B	C	were bombing shipping in the North Sea and threatening to	invade	England on 18 July . Churchill had made a speech after the
19	AE4	W_non_ac_humanities_arts	A	B	C	d'Oysel and told him " that in no wise would they	invade	England " . This was not , of course , just a Protestant
20	CS5	W_non_ac_humanities_arts	A	B	C	South America , which convinced Philip that he should try to	invade	England . By the 1590s it was clear that the Netherlands would
21	EE5	W_biography	A	B	C	's Imperial Army who had died of disease while waiting to	invade	England . In the middle of the wood was an obelisk commemorating
22	AS7	W_fict_prose	A	B	C	Scottish army gathered at Caddonlea near Selkirk to	invade	England . They were ignominiously defeated by the Earl of Surrey
23	CEP	W_newsp_other_sports	A	B	C	to the Champions League , begin their quest to successfully	invade	Europe against Marseille at Ibrox tonight . But their captain ,
24	C95	W_pop_lore	A	B	C	disease -- Myxobacteria (Cytophaga) These organisms will	invade	fish tissues and multiply when conditions suit them .
25	K5J	W_newsp_other_social	A	B	C	the spectators who take the law into their own hands and	invade	football pitches . At the weekend , the world indoor athletics
26	FDF	W_ac_humanities_arts	A	B	C	need of repair by the time that the English began to	invade	France . But by galvanising men into action , the English

***Корпусний аналіз вирізняється низкою характерних ознак*:**

1) емпіричний підхід до аналізу мовних даних (досліджуються реальні моделі мовної реалізації у природних текстах);

2) використання великих за обсягом, структурованих колекцій природних текстів (корпусів) як основи для аналізу;

3) широке залучення комп'ютерних технологій для дослідження лінгвального матеріалу;

4) застосування квалітативних і квантитативних аналітичних методик, з суттєвою перевагою останніх (вивчення частоти вживання лінгвістичних одиниць, статистичні дослідження сполучуваності і т. ін.).

Спираючись головним чином на реальний «живий» мовний матеріал, а не на мовну інтуїцію та інтроспекцію, корпусні дослідження дозволяють абстрагуватися від суб’єктивності дослідника і наблизитися до об’єктивного вивчення мови.

Корпусні розвідки переорієнтовують традиційний підхід до вивчення мови, а результати аналізу даних корпусу сприяють переоцінці низки лінгвістичних теорій [MacEnery, Hardie, 2012: 1].

Напрями корпусного мовознавства

Перший напрям зосереджений на розробці проблем, що стосуються теорії та практики створення корпусів.

Другий напрям спрямований на дослідження саме лінгвістичних корпусів, тобто вивчення мови за допомогою корпусних методів

Двовекторність корпусної лінгвістики зумовлюється подвійною природою *об'єкта* її дослідження – текстового корпусу, який, з одного боку, виступає в якості вихідного мовленнєвого матеріалу для корпусної лінгвістики, а з іншого, є результатом діяльності цього мовознавчого напряму.

Предметом **корпусної лінгвістики виступають теоретичні основи і практичні механізми створення та експлуатації мовних корпусів.**

Першочерговою *метою* КЛ є об'єктивний лінгвістичний опис мовної системи, причому до цього опису корпусна лінгвістика підходить від вивчення конкретної людської комунікації.

У якості другорядної цілі розглядається вироблення особливого способу відображення мовного матеріалу в корпусі текстів.

Теоретичним підґрунтям корпусної лінгвістики є структуралізм, який декларує примат реального тексту в лінгвістичному дослідженні.

Для корпусних розвідок головним є постулат, що мова як об'єкт дослідження може бути вивчена лише у формі писемних та усних текстів [Демьска 2010: 6].

Дослідницька програма корпусної лінгвістики

1) КЛ є суто емпіричною дисципліною й при аналізі лінгвального матеріалу покладається на реальне функціонування мови з метою встановлення правил та вивчення особливостей продукування мови людиною, на відміну від тих досліджень, які опираються на вигадані приклади чи інтроспекцію.

2) **Застосування комп’ютерів дозволяє миттєво обробити величезний обсяг мовного матеріалу і відібрати всі можливі у конкретному корпусі приклади вживання необхідних для аналізу одиниць. У розпорядження лінгвіста надаються об’єктивні кількісні дані, забезпечуючи досягнення більш ґрунтовних та переконливих висновків.**

3) Корпусна лінгвістика дозволяє вченим підтвердити або спростувати гіпотези про функціонування мови, а також окреслити нові напрями дослідження, які до застосування корпусних методів не попадали до фокусу уваги дослідників.

2. Корпусна лінгвістика в системі мовознавчих наук

1) методологія аналізу мови
2) самостійна дисципліна прикладного мовознавства

Корпусна лінгвістика має принаймні дві ознаки, що дають їй підставу претендувати на статус самостійної дисципліни:

1) характер аналізованого словесного матеріалу;

2) специфіка інструментарію [Захаров, Богданова, 2011: 9].

Корпус – це не просто новий і потужний інструмент: за використанням корпусу стоїть певна ідеологія, основні тенденції якої зародилися ще в класичній філології XIX століття, але значно інтенсифікувалися в останні десятиліття [Плунгян, 2008: 7–20.].

Головними пріоритетами цієї ідеології є:

1. **увага не до слова чи речення, а до *тексту* (дискурсу), тобто до реального інструменту комунікації в цілому, а не до його окремих фрагментів;**

2. **увага до *квантитативного компонента мови*, тобто врахування в першу чергу більш частотних елементів порівняно з менш частотними, визнання квантитативних відношень суттєвим фактором у мовній еволюції і структурі мовних правил;**

3. увага до *синхронічної варіативності мови*, тобто визнання того факту, що не існує єдиної жорсткої системи засобів вираження змісту, а існують її різні реалізації, в тому числі залежні від психологічних, біологічних і соціальних факторів;

4. увага до *діахронічної варіативності мови*, тобто визнання того факту, що мова постійно змінюється у часі і повністю відволіктися від цієї нестабільності не можливо, в кожен момент часу в мові співіснують «прогресивні» і «консервативні» ділянки;

5. зміна відношення до поняття мовної норми і мовної правильності, тобто межа між «помилкою» та «маргінальним варіантом» визнається більш рухомою та хиткою [Плунгян, 2008: 7–20.].

Корпусна лінгвістика як емпіричний мовознавчий напрям суттєво відрізняється від традиційної лінгвістики підходами та методами вивчення мовного матеріалу

Корпусна лінгвістика	Традиційна лінгвістика
Основна увага – вивчення мовлення	Основна увага – вивчення мови
Мета – опис мови у тому вигляді, як вона проявила себе в мовленні, представленого у вигляді спеціально відібраного корпусу текстів	Мета – опис та пояснення мови

У своїх дослідженнях спирається на дані корпусу тексту	У своїх дослідженнях йде від теорії до її пояснення і підтвердження у фактах мовлення
Надає перевагу квантитативним методам	Надає перевагу квалітативним методам
Вбачає себе частиною традицій, що базується на емпіричних методах	Вбачає себе частиною традицій, що базується на раціоналістичних методах
Текст розглядається як певна фізична сутність	Текст розглядається як певна абстракція
Укладення граматики конкретних мов	Вивчає мовні універсалії
Основна увага приділяється формі	Головна увага – не лише формі, але і змісту

Проводиться робота з лінгвістичними даними (слововживаннями) у тому вигляді, в якому вони зустрілися в контексті	Надає перевагу штучним прикладам з ізольованих від тексту слововживань
Надає перевагу індуктивним методам обробки емпіричного словесного матеріалу, вважає їх суттю наукового методу	Надає перевагу дедуктивним методам обробки емпіричного словесного матеріалу
Вірить у наукові відкриття, базовані на обробці емпіричних даних	Вірить у відкриття, базовані на процедурах, оцінках, порівняннях і т.ін., як результат багатовікових досліджень

ВИСНОВОК:

корпусні студії змінюють пріоритети сучасних лінгвістичних досліджень і демонструють виразну переорієнтацію об’єкта дослідження з «системи» на «узус», з «мови» на «мовлення».

Традиційне мовознавство вивчало *можливість (possibility)* або неможливість якого-небудь лінгвістичного явища, а корпусна лінгвістика додатково вивчає й *імовірність (probability)* лінгвістичних явищ.

Корпусна vs комп'ютерна лінгвістика

1) Функція мови
2) Застосування комп'ютерних інструментів
3) Інтелектуальна інтерпретація даних
4) Комп'ютерні програми

3. Типологія досліджень у царині корпусного мовознавства

Сьогоднішня корпусна лінгвістика – це гетерогенна область дослідження мови, всередині якої виокремлюються окремі піднапрями, що різняться підходами до конструкції, експлуатації корпусів та аналізу корпусних даних. В основі виділення цих піднапрямів знаходяться такі параметри [McEnery, Hardie 2012: 3-21]:

- **формат представлення текстів у корпусі (*mode of communication*);**
- **корпуснобазовані (*corpus-based*) vs. корпуснокеровані (*corpus-driven*) дослідження;**
- **режим накопичення даних у корпусі (*data collection regimes*);**
- **використання анотованих (*annotated*) / неанотованих (*unannotated*) корпусів;**
- **повне врахування (*total accountability*) vs відбір даних (*data selection*);**
- **багатомовні (*multilingual*) vs одномовні (*monolingual*) корпуси.**

Критика корпусних досліджень

ЛЕКЦІЯ 2

КОРПУСНІ СТУДІЇ: ІСТОРИЧНА ПЕРСПЕКТИВА ТА СУЧАСНИЙ СТАН

План

1. Історія становлення корпусної лінгвістики: від паперових конкордансів і картотек до перших електронних корпусів
2. Корпусна лінгвістика з 60-х років XX ст. до пост 2000-х
3. Корпусні дослідження в Україні

Етап 1 **(середина 60-х – початок 80-х років XX століття) – період набуття знань про організацію та підтримку корпусів до 1 млн. слів, характеризується відсутністю матеріалів в електронному форматі та потребою набору текстів вручну.**

Eman 2 **(1980–2000 рр.) поділяється на два періоди :**

1. **1980-ті роки відзначилися появою сканерів, коли навіть із примітивним сканером укладалися корпуси у 20 млн. слововживань;**
2. **1990-ті роки ознаменовані розширенням можливостей комп'ютерного набору, що полегшило доступ до великих за обсягом текстових матеріалів в електронному форматі і сприяло значному збільшенню розмірів корпусів.**

Eman 3 **(з початку 2000-го року і по сьогоднішній день) – це період електронних (віртуальних) текстів, які ніколи не мали матеріальної форми, що надає величезні можливості для створення корпусів будь-якого необмеженого розміру [Tognini-Bonelli, 2010: 16-17].**

У. МакЕнері та А. Вільсон

- **Перший період – це стадія ранньої корпусної лінгвістики (1910–1960-ті рр.), коли відбувається формування теоретичного підґрунтя та прагматичних передумов виникнення напряму й створення текстових зібрань для лінгвістичного дослідження переважно на паперових носіях.**
- **Другий період (починається з 1960 рр.) характеризується інтенсивним піднесенням корпусних студій і безпосередньо пов'язаний із значним розвитком комп’ютерних технологій.**

До 1990-х у корпусних дослідженнях чітко окреслилися три напрями теорії та практики:

1) побудова електронних текстових корпусів;

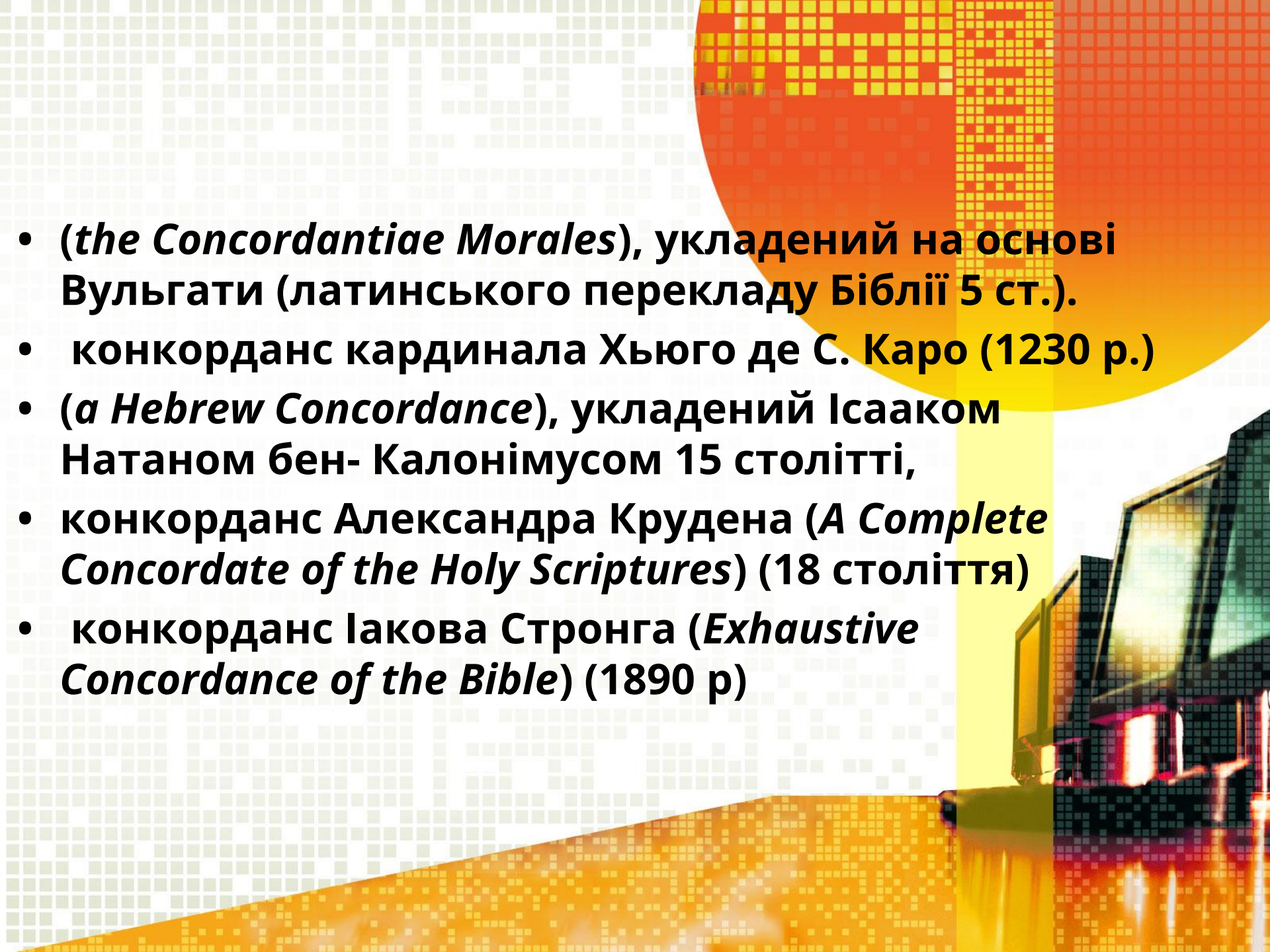
2) програмне опрацювання текстових корпусів;

3) екстрагування, аналізу й опису корпусних даних [Демська ст. 10].

Доелектронні корпуси. Конкордансі Біблії

Конкорданс – це алфавітний список усіх вжитих у певному тексті/текстах слів у їх контексті.

- (*the Concordantiae Morales*), укладений на основі Вульгати (латинського перекладу Біблії 5 ст.).
- конкорданс кардинала Хьюго де С. Каро (1230 р.)
- (*a Hebrew Concordance*), укладений Ісааком Натаном бен- Калонімусом 15 столітті,
- конкорданс Александра Крудена (*A Complete Concordate of the Holy Scriptures*) (18 століття)
- конкорданс Іакова Стронга (*Exhaustive Concordance of the Bible*) (1890 р)



Конкорданси літературних творів

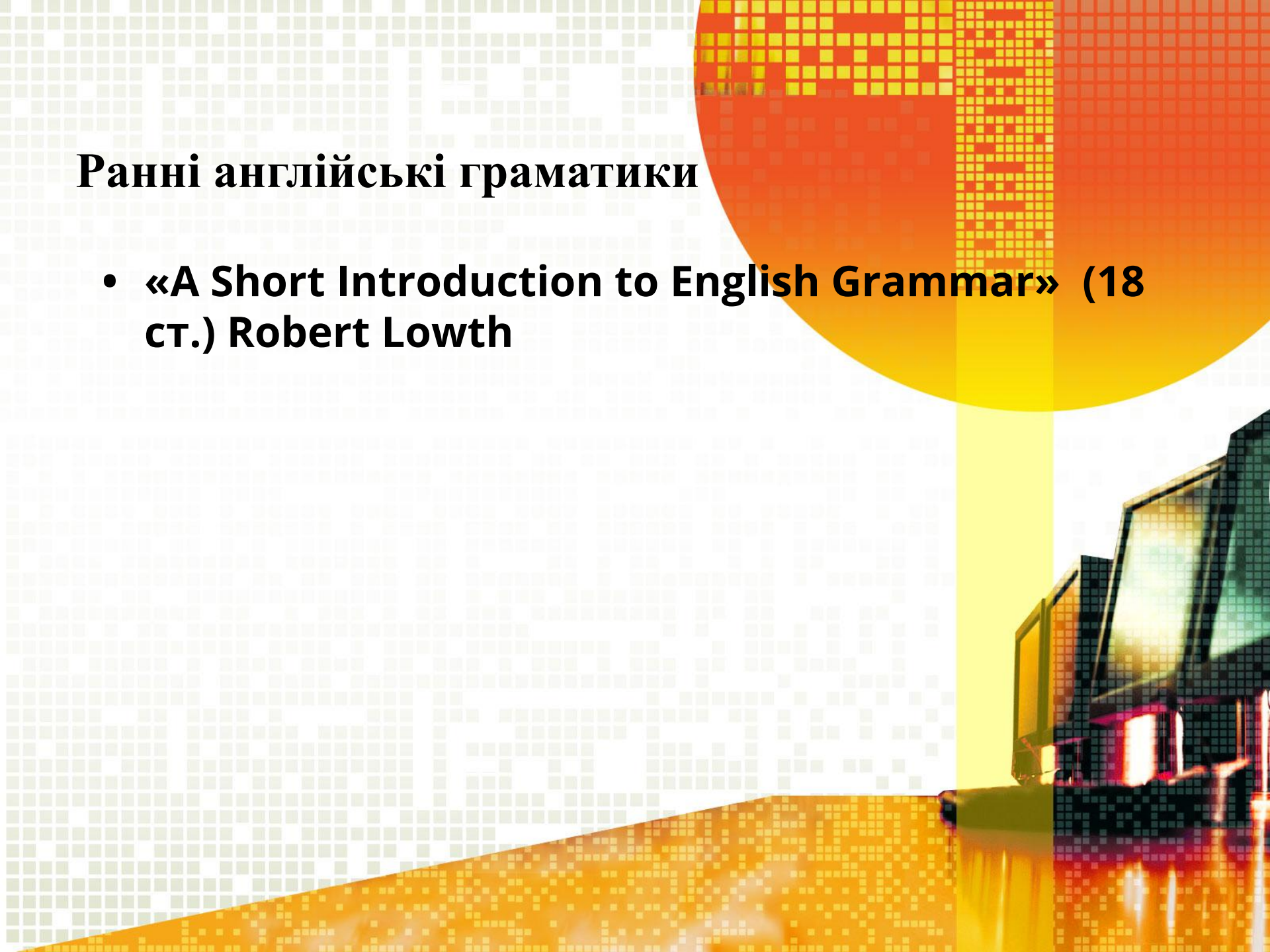
- конкорданс праць У. Шекспіра Ендрю Бекета (*A Concordance of Shakespeare*) (1787 р.),
- конкорданс праць Дж. Чосера, що був укладений у 1871 році, опублікований у 1927 році.

Корпуси для укладання ранніх граматик

- **граматика Паніні 4 столітті до н.е.**
- **“Неграматичні слова” Аристона Алекасандрійського (1 століття н.е.)**

Ранні англійські граматики

- «A Short Introduction to English Grammar» (18 ст.) Robert Lowth



- **О.Єсперсен (1909-1949) «A Modern English Grammar on Historical Principles»**

- ***It is impossible for me to put even a remotely accurate number on the quantity of slips I have had or still have: a lot of them have been printed in my books, particularly the four volumes of Modern English Grammar, but at least just as many were scrapped when the books were being drafted, and I still have a considerable number of drawers filled with unused material. I think a total of 3-400,000 will hardly be an exaggeration [Jespersen 1938: 213-215; translation by D. Stoner].***

- **George Curme, Hendrik Poutsma, and Charles Fries**

Укладання словників

- Словник Самуеля Джонсона (1755)

- **Джонсон зібрав 150,000 ілюстративних цитат для 40,000 заголовних слів словника, а читачі Oxford English Dictionary зібрали 5 млн. цитат для ілюстрацій 400,000 слів.**

- **Найважливішим та найвпливовішим доелектронним корпусом вважається The Survey of English Usage, укладений Рендольфом Квірком у 1959 р. в University College London.**

- **http://www.ucl.ac.uk/english–usage**

Корпусна лінгвістика у 60-ті р. ХХст.

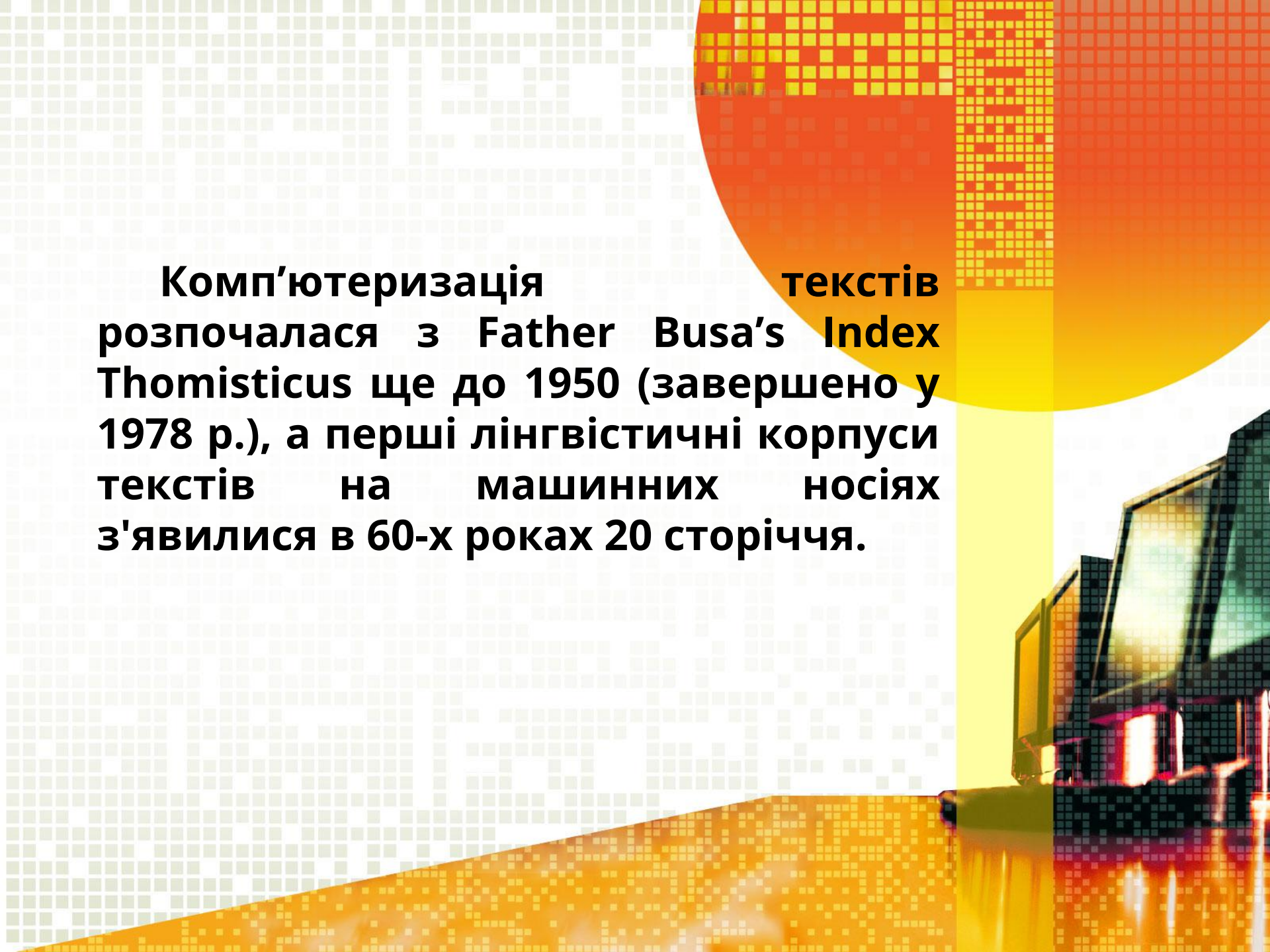
- Переважна кількість досліджень у царині сучасної корпусної лінгвістики розпочиналася на матеріалі англійської мови.

- **Корпусні студії були неоднозначно сприйняті у науковій спільноті та зазнали суттєвої критики від засновника генеративізму Н. Хомського.**

- **Дослідник назвав корпусний спосіб накопичення мовних даних неадекватним і хибним для опису породжувальної здатності природної мови, оскільки лише інтуїція мовця може замінити корпус і стати джерелом мовного матеріалу**

Ідея створення корпусу (вже у сучасному його розумінні) зародилася у 60-х роках 20 століття

Комп’ютеризація текстів розпочалася з Father Busa’s Index Thomisticus ще до 1950 (завершено у 1978 р.), а перші лінгвістичні корпуси текстів на машинних носіях з'явилися в 60-х роках 20 сторіччя.



Корпуси першого покоління

Перший мільйонний корпус текстів на машинному носії було укладено у 1963 р. в Браунівському університеті (США) (the Brown Corpus).

- **автори У. Френсис і Г. Кучера**
- **дослідження лінгвістичних особливостей американського варіанту англійської мови**
- **містив 500 текстових уривків обсягом по 2 000 слововживань загальним обсягом біля 1 млн. слів.**

Корпус супроводжувався значною кількістю матеріалів його первинної статистичної обробки — частотний і алфавітно-частотний словник, різноманітні статистичні розподіли.

У. Френсіс та Г. Кучера ставили собі мету представити корпус текстів, що відповідав ясним і чітким критеріям відбору.

Укладачами враховувалися такі характеристики, як:

1. ***походження і склад тексту*** **(автор повинен був бути уродженим носієм американського варіанту англійської мови, діалогічне мовлення повинно було займати менше половини всього обсягу тексту);**
2. ***часова віднесеність*** **(всі відібрані до корпусу тексти були вперше опубліковані у 1961 році);**
3. ***збалансоване представлення різних жанрів*****;**
4. ***доступність для комп'ютерної обробки*** **(спеціальні помітки для передачі графічних особливостей тексту і т. п.).**

- **Поява Браунівського корпуса викликала загальний інтерес у колі лінгвістів і жваві дискусії.**

- **Браунівський корпус швидко перетворився у популярний об’єкт дослідження і навіть в певний стандарт для створення інших аналогічних корпусів.**

Поступово в процесі його використання вчені дійшли до розуміння того, що провести певні порівняння і виявити конкретні закономірності можливо лише шляхом аналізу значних за розміром масивів текстів, які організовані за визначеними правилами. Так почали проводитися нові дослідження мови вже на більш високому і надійному рівні в межах нового напряму в лінгвістиці, яким стала корпусна лінгвістика.

- **Услід за Браунівським корпусом з’явилися британський аналог Браунівського корпусу – Ланкастерсько-Осло-Бергенський корпус (Lancaster-Oslo-Bergen Corpus)**

- **Створення Браунівського та Ланкастерського корпусів дало можливість проводити різноаспектні філологічні порівняння двох варіантів англійської мови (американського і британського) на текстах різних жанрів**

За форматом Браунівського та Ланкастерсько-Осло-Бергенського корпусів з деякими модифікаціями було укладено низку інших корпусів, серед яких the Kolhapur Corpus of Indian English, the Wellington Corpus of Written New Zealand English, the Australian Corpus of English, the Corpus of English-Canadian Writing, the Standard Corpus of Present-day English Language Usage, the London-Lund Corpus (LLC)

- **70-ті роки 20 століття були періодом уповільнення темпів корпусних досліджень.**
- **у 80-ті роки 20 століття у світі було здійснено декілька спроб створити корпуси обсягом більше 1 млн**

Корпуси другого покоління

- Перший мега-корпус, що задав новий стандарт для представницьких корпусів – Британський національний корпус (British National Corpus). http://www.natcorp.ox.ac.uk/

- **Цей корпус характеризується обсягом 100 млн. слів, використанням повних текстів, а не вибірок з текстів, підкорпусом усного мовлення (10 млн. слів), наявністю частиномовної розмітки та доступом через Інтернет. Для корпусу використовувалася детальна класифікація документів за декількома параметрами: вид мовлення (писемне, усне приватне і усне публічне), для писемного за тематикою, типом видання (книги, періодика, машинописні тексти і т. п.), параметром утворення очікуваної аудиторії (високий, середній чи довільний) та складністю мови**

Укладачі BNC для порівняння спробували представити корпус у вигляді звичайної книжкової продукції і одержали вражаючі показники. Якщо видрукувати корпус на тонкому папері з розрахунку 400 слів на сторінку, то весь його обсяг у друкованому вигляді займатиме простір близько 10 м^2. Для того, щоб прочитати цю продукцію зі швидкістю 150 слів на хвилину, витрачаючи на це 8 годин щодня, знадобилося б 4 роки [Карпіловська 2006: 76]

- **За заданим Британським національним корпусом стандартом були укладені представницькі корпуси багатьох європейських мов. За цією моделлю були створені національні корпуси іспанської, італійської, хорватської, чеської мов.**

- **Подібний проект Банк англійської мови (the Bank of English) розпочався у 1980-і рр. У 1989 році його обсяг був 20 млн. слів, а у 2012 – 650 млн. слів.**

- **Банк англійської мови – це так званий моніторинговий корпус, що покликаний відслідковувати мовні зміни шляхом регулярного поповнення новими текстами та порівняння частотних параметрів, наприклад, таких, як зміна частоти слів та граматичних конструкцій, поява нових слів і т.ін. Він охоплює англійське писемне та усне мовлення, а також різні територіальні варіанти англійської мови.**

- **Банк англійської мови та Британський національний корпус мали потенційну підтримку від видавців, що використовували корпуси для укладання словників і граматик. Такими ж корпусами є Кембриджський та Лонгманівський корпуси, що є закритими для вільного доступу і використовуються лише авторами та укладачами навчальних матеріалів видавництв.**

- **Інтернаціональний корпус англійської мови (the International Corpus of English)**
- **the American National Corpus**
- **Машинний Фонд російської мови**

- **У 1992 році була створена організація Європейська корпусна ініціатива (ECI), метою якої були об’єднання і координація зусиль лінгвістів різних країн, що працюють над створенням корпусів текстів на інших, крім англійської, мовах. Під її егідою було створено біля 50 корпусів текстів (кожен обсягом від 12 тисяч до 5 млн. слів) на європейських мовах.**

- **Сучасний розвиток корпусної лінгвістики (пост 2000-і роки) дуже бурхливий, що підтверджується величезною кількістю нових досліджень у галузі.**

- **дослідження у галузі лексичної граматики [Stubbs 1996; Hunston, Francis 2000; Renouf 2001; Nesselhauf 2005; Exploring the Lexis-Grammar Interface 2009],**
- **лексикографії та навчання мові [McEnery, Kifle 2002, Altenberg, Granger 2002; McEnery, Xiao 2004, Максимів 2008],**
- **когнітивної лінгвістики [Corpora in Cognitive Linguistics 2006; Gilquin 2003; Gries 2003; Gries, Stefanowitch 2004; Schmidt 2000; Schonefeld 1999],**
- **прагматики та дискурс-аналізу [Aijmer and Stentström 2004; Archer 2005; Baker 2005; Baker, McEnery 2005; Hardt-Mautner 1995; Koller, Mautner 2004; McEnery 2005; Orpin 2005; Partington et al. 2004; Vivanco 2005; Wang 2005],**

- **стилістики [Burrows 2002; Charteris-Black 2004; Corpus-Based Approaches to Metaphor and Metonymy 2006; Deignan 2005; Semino and Short 2004; Stubbs 2005],**
- **перекладознавства [Malmkjær 1998; Zanettin 1998; Incorporating Corpora. The Linguist and the Translator 2008].**
- **Корпусно-базовані дослідження відбуваються для вивчення значення слова [Partington 2004], фразеології [Hunston 2001, Лозинська 2009], синтаксичних властивостей граматичних структур [Duffley 2003], дистрибуції граматичних категорій [Biber 2001] і т.ін.**

- **Найновіші досягнення в царині корпусного мовознавства друкуються у визнаних міжнародних наукових журналах:**
- **Corpus (2001–) (Nice: Laboratoire "Bases, Corpus, Langage«),**
- **Corpus Linguistics and Linguistic Theory (2005–) (Berlin – New York: Mouton De Gruyter)16;**
- **ICAME Journal, Journal of the International Computer Archive of Modern English (1987–) (Bergen: Norwegian Computer Centre for the Humanities)17;**
- **International Journal of Corpus Linguistics (1996–) (Amsterdam: John Benjamins) 18;**
- **Language Resources and Evaluation (2005–) (Dordrecht: Springer)19;**
- **Literary and Linguistic Computing (1986–) (Oxford: Oxford University Press)**

- **У цей час корпуси створені для багатьох мов світу (див. веб сайт Дейвіда Лі, http://www.uow.edu.au/~dlee/CBLLinks.htm)**

- **Ч. Філмор [Fillmore 1992: 35] зазначив, що навіть значні за обсягом корпуси не в змозі відобразити все можливе у мові, натомість і невеликі за обсягом корпуси можуть надати інформацію, яку б нереально було отримати, не звертаючись до корпусних даних.**

3. Корпусні дослідження в Україні

- **Український національний лінгвістичний корпус (УНЛК) - 100 млн. слововживань**

Корпус текстів природної мови.

- Поняття “корпус текстів”
- Типологія корпусів.
- Типи корпусної розмітки.

- **Доцільність створення й використання корпусів визначається такими передумовами:**
- **1) досить великий (репрезентативний) обсяг корпусу гарантує типовість даних і забезпечує повноту представлення всього спектру мовних явищ;**
- **2) дані різного типу перебувають у корпусі у своїй природній контекстній формі, що створює можливість їх всебічного й об'єктивного вивчення;**
- **3) одного разу створений і підготовлений масив даних може використовуватися багаторазово, багатьма дослідниками й у різних цілях [Захаров, Богданова 2011: 8].**

Підходи до трактування поняття “корпус”

- **корпус – це організована певним чином словесна єдність, елементами якої є цілі тексти чи спеціальним чином відібрані уривки з текстів, що доступні для лінгвістичного аналізу [Meyer 2004: xi];**

- **корпус – це зібрання текстів, яке вважається репрезентативним стосовно даної мови, діалекту або іншої ділянки мови й призначене для використання в лінгвістичних дослідженнях [Francis 1991];**

- **корпус – це певне зібрання текстів, в основі яких лежить логічний задум, логічна ідея, що об’єднує ці тексти. Логічна ідея втілюється в правилах організації текстів в корпус, алгоритмі і програмі аналізу корпусу текстів та в пов’язаних з цим ідеологією та методологією. Корпус є четвертою фактурою мовлення (тексти на машинному носії) [Рыков27];**

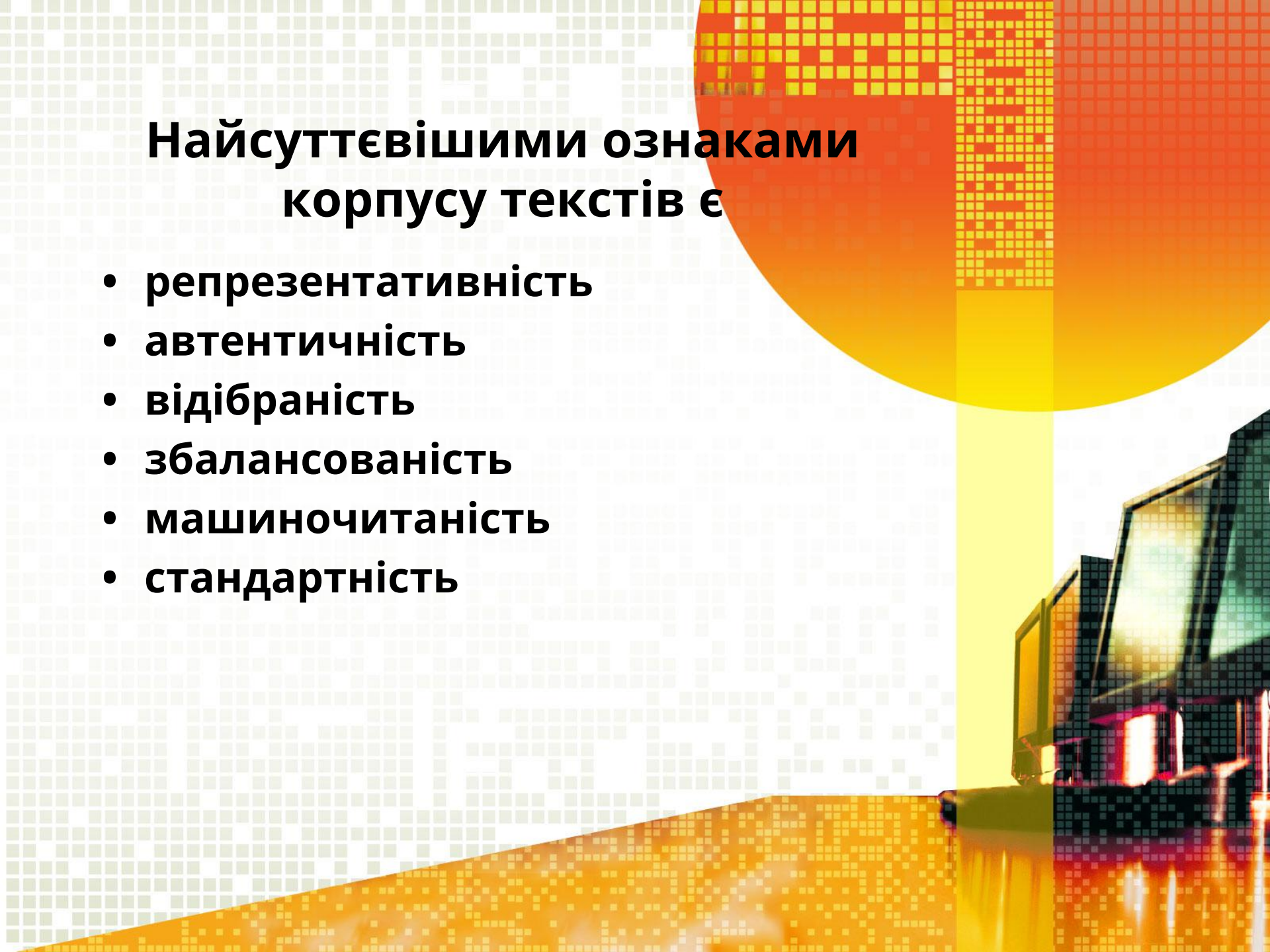
- **корпус – це машиночитане, стандартно організоване зібрання репрезентативних для певної мови, діалекту або іншої підмножин(и) мов(и) писемних або усних текстів, призначених для лінгвістичного аналізу й опису, відібраних і впорядкованих згідно з експліцитними екстра- та інтралінгвальними критеріями [Демська-Кульчицька 2005].**

Комп’ютерний корпус текстів характеризується такими ознаками як

- логічна єдність задуму;
- кінцевий розмір;
- обов’язкове його розміщення на машинному носії;
- стандартне представлення чи розмітка словесного матеріалу в корпусі для зручності його програмної обробки.

Найсуттєвішими ознаками корпусу текстів є

- репрезентативність
- автентичність
- відібраність
- збалансованість
- машиночитаність
- стандартність



- **У типології корпусів В.В. Рикова виділяються такі типи28:**
- **1. За ступенем організації й структурованості:**
- **□ електронний архів – це тексти на електронному носії, але форма їх представлення на машинному носії не стандартизована й не уніфікована;**
- **□ електронна бібліотека – тексти тут представлені однорідним і стандартизованим способом;**
- **□ корпус текстів – форма стандартизована й уніфікована, тексти призначені для відображення частини лінгвістичної реальності;**
- **□ субкорпус – це деяка автономна частина корпуса..**

- **2. За хронологічною ознакою:**
- **□ синхронічний;**
- **□ моніторинговий (відслідковує поточний стан мови**
- **□ діахронічний.**
- **3. За індексацією:**
- **□ простий;**
- **□ анотований.**

- **4. За мовою:**
- **□ одномовний;**
- **□ двомовний;**
- **□ багатомовний.**
- **5. За способом застосування й використання корпусу:**
- **□ дослідницький;**
- **□ ілюстративний;**
- **□ паралельний.**
- **6. За способом існування корпусу:**
- **□ динамічний;**
- **□ статичний**

Класифікація корпусів (за О. Демською-Кульчицькою)

- **За типом подання тексту:**

повнотекстові - фрагментарні

- **За стратегією побудови і використання:**

дослідницькі - ілюстративні

- **за типом реалізації мовної системи:**

усні - писемні - змішані

- **За способом подання мовного матеріалу:**

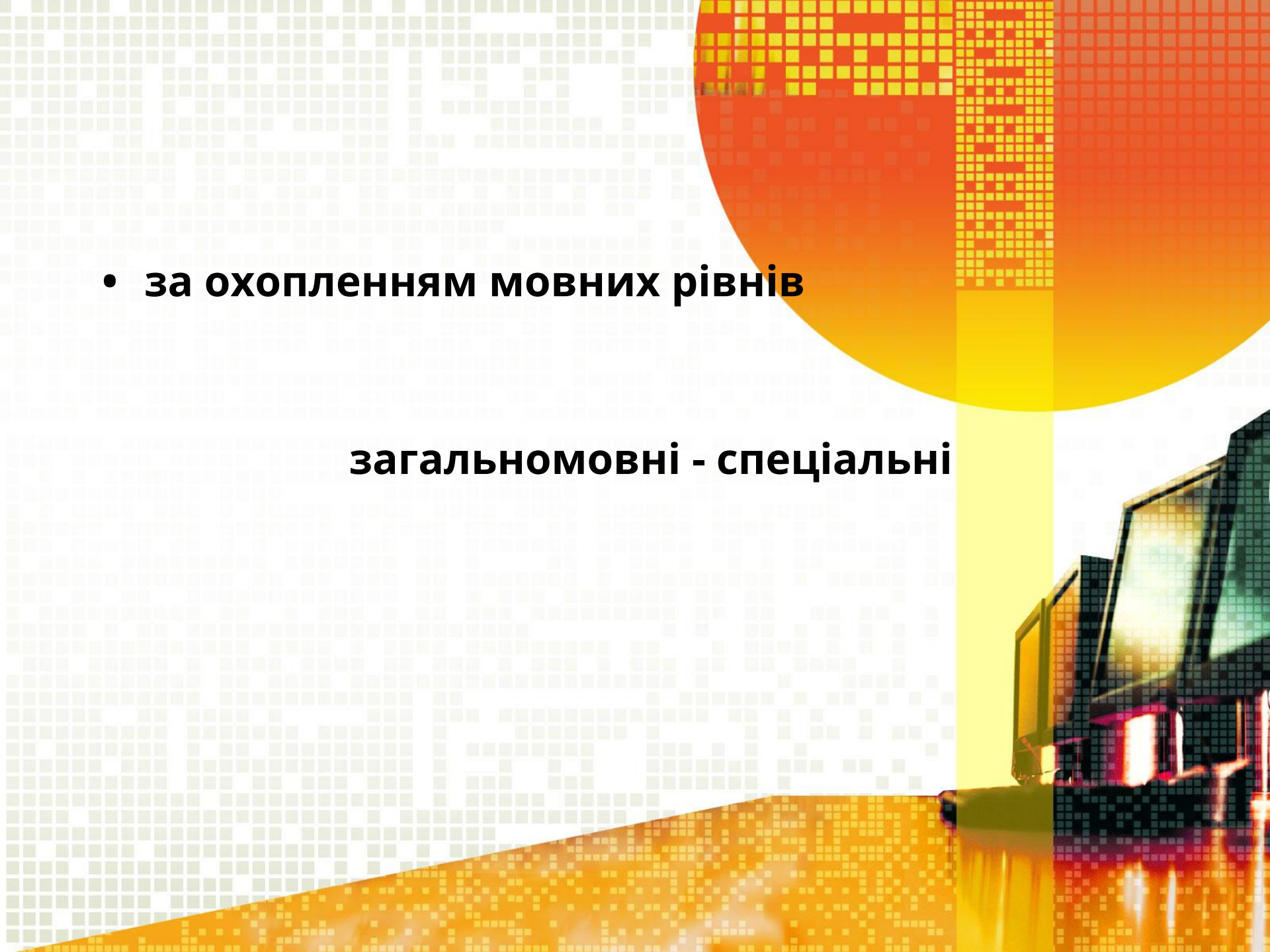
динамічні - статичні

- **За хронологічними параметрами:**

діахронні - синхронні

- **за охопленням мовних рівнів**

загальномовні - спеціальні



- **за кількістю мов**

одномовні - багатомовні

- **За типом кореляції мов:**

паралельні - порівняльні

- **за обсягом**

малі-середні-великі-надвеликі

- **За типом кодування**

неанотовані - анотовані

Національний корпус

- British National Corpus (обсяг 100 млн. слововживань), the American National Corpus (22 млн.) , the PELCRA Referenc Corpus of Polish Corpus (100 млн.), the Czech National Corpus (більше 100 млн.), the Hungarian National Corpus (187,6 млн.), the Hellenic National Corpus (корпус сучасної грецької мови, загальним обсягом 47 млн. слововживань), the DWDS corpus (обсяг 100 млн. слововживань), the Slovak National Corpus (339 млн.), the Modern Chinese Language Corpus (100 млн. знаків) та інші

Спеціалізований корпус

- the Guangzhou Petroleum English Corpus
- The Michigan Corpus of Academic Spoken English (MICASE)
- the Epistolary Corpus of Victorian Women Writers' Letters, the Shakespeare Corpus, Корпус словаря языка Достоевского)

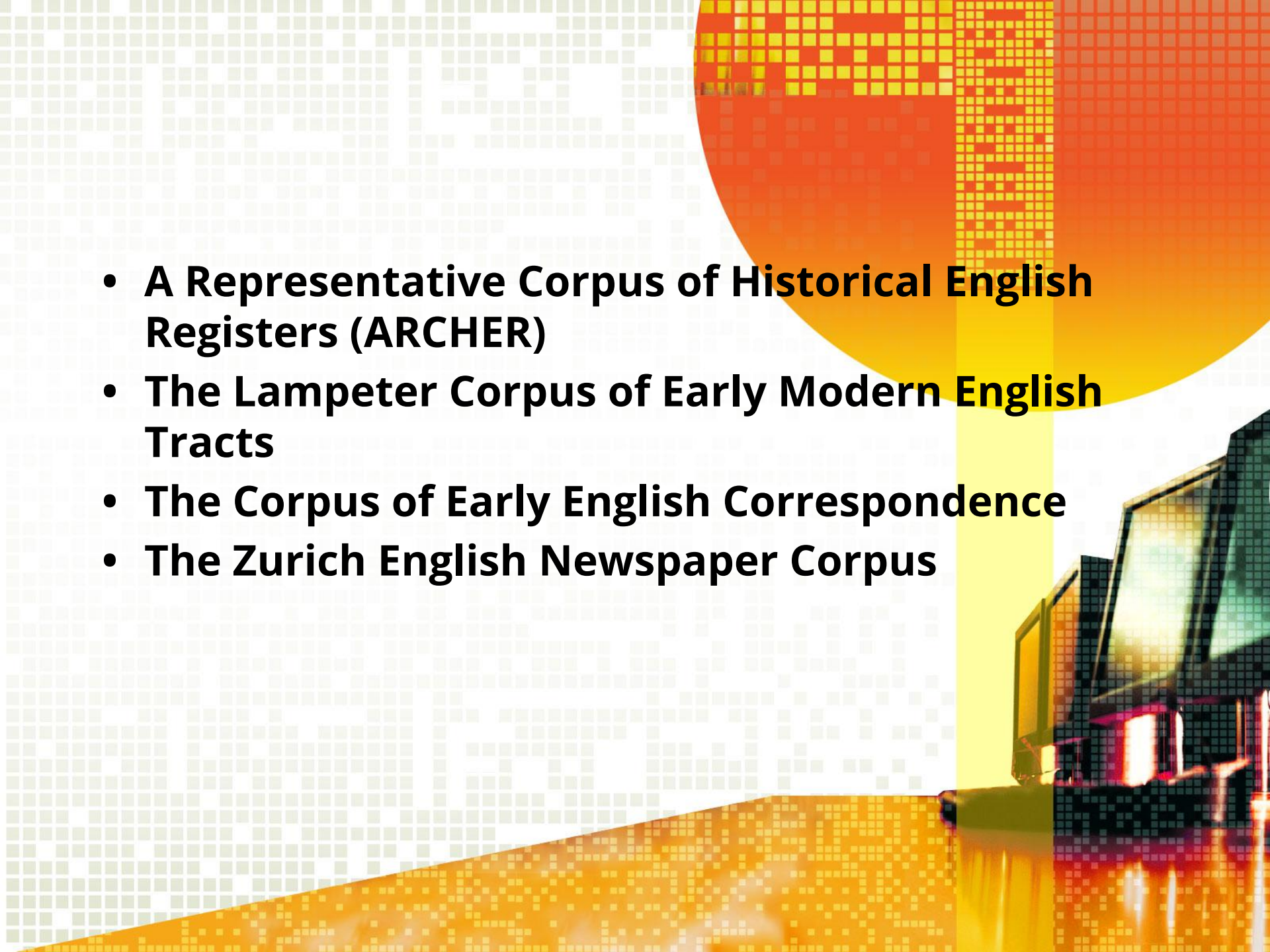
Учнівські корпуси

- the Longman Learners’ Corpus
- the Cambridge Learner Corpus,
- the International Corpus of Learner English
- the Hong Kong University of Science and Technology Learner Corpus

Історичні корпуси

- **Helsinki Corpus of English Texts (the Brooklyn-Geneva-Amsterdam-Helsinki Corpus of Old English, the Penn-Helsinki Parsed Corpus of Middle English (1150-1500 pp., 1,2 млн. слововживань), the Penn-Helsinki Parsed Corpus of Early Modern English (1500-1700 pp., 1,7 млн. слововживань), the Penn Parsed Corpus of Modern British English (1700-1914 pp., 1 млн. слововживань).**

- A Representative Corpus of Historical English Registers (ARCHER)
- The Lampeter Corpus of Early Modern English Tracts
- The Corpus of Early English Correspondence
- The Zurich English Newspaper Corpus



Корпуси писемного\усного мовлення

- the Australian Corpus of English репрезентує австралійський писемний варіант англійської мови (1986-)
- The Wellington Corpus of Written NZ English (WWC) представляє новозеландський писемний варіант англійської мови (1986-1990 pp.)
- the Kolhapur Corpus відображає індійський писемний варіант англійської мови (1978-)

- **the LondonLund Corpus (LLC), the Lancaster/IBM Spoken English Corpus (SEC), the Cambridge and Nottingham Corpus of Discourse in English (CANCO DE), the Santa Barbara Corpus of Spoken American English (SBCSAE) та the Wellington Corpus of Spoken New Zealand English (WSC)**

Поняття корпусної розмітки

1. Типи корпусної розмітки
2. Вимоги до розмітки

- **Лінгвістичний корпус за визначенням є такою колекцією природно мовних текстів, де здійснено розмітку (маркування) хоча б за одним лінгвістичним параметром. Ця ознака є такою, що вирізняє лінгвістичний корпус з-поміж великого числа інших лінгвістичних інформаційно-інструментальних систем, баз даних та знань [Корпусна лінгвістика 2005: 33].**

- **Процес розмітки (*tagging, annotation*) полягає в приписуванні текстам і їх компонентам спеціальних міток (*tag, tags*):**

- ***зовнішніх, екстралінгвістичних*** **(відомості про автора й відомості про текст: автор, назва, рік і місце видання, жанр, тематика; відомості про автора можуть включати не тільки його ім'я, але також вік, стать, роки життя й багато чого іншого (це кодування інформації має назву *метарозмітка*);**
- ***структурних*** **(розділ, абзац, речення, словоформа);**
- **власне *лінгвістичних*, що описують лексичні, граматичні та інші характеристики елементів тексту.**

- **анотація (*annotation*) :: структурне маркування (*markup*)**

- **«процес анотування корпусних даних – це додавання інтерпретованої, лінгвістичної інформації до електронного корпусу усного чи/або писемного мовлення» [Leech 1997: 2].**

- **Маркування надає відносно об’єктивну верифіковану інформацію про частини корпусу та структуру кожного тексту [McEnery, Xiao, Tono 2006: 29].**

структурна анотація (corpus markup)

- **Ч. Меєр використовує цю терміносполуку на позначення і структури тексту, і зовнішньої стосовно нього інформації (його бібліографічний опис, дані про мовців тощо) [Meyer 2002: 81]**
- **Ґ. Астон і Л. Бернард: “…корисно вказувати межі глав, розділів, абзаців, речень, і т. д., а також особливу роль заголовків, переліків, приміток, посилань, супровідних підписів, покликів та ін.” [Aston , Burnard 1998: 24].**

- **Під елементами універсальної структури тексту розуміються <head> (заголовок), (частина, розділ), <p> (абзац), <s> (речення), <epigraph> (епіграф), <dateline> (дата), <note> (примітка), <said> (пряма мова), <dedication> (присвята), <l> (рядок, у вірші), <abbr> (скорочення), <num> (число) та ін.**

- **Отже, структурою тексту вважаємо такі його елементи, як назва, розділ, підрозділ, рубрика, присвята, епіграф, поклик, цитата, вживання алфавітів інших писемних систем, цифр тощо. Структурне анотування – це виділення структурних елементів тексту за допомогою певної мови маркування; сукупність маркерів-вказівок на елементи зовнішньої будови тексту.**

лінгвістична анотація

Під лінгвістичною анотацією у корпусній лінгвістиці традиційно розуміють:

а) довільну лінгвістичну інформацію про лінгвально релевантні одиниці текстових даних, поданих через формальний код;

б) практику введення формалізованої лінгвістичної інформації в електронний текст;

в) наявність такої інформації у тексті [Демська-Кульчицька 2004: 26].

- **Морфологічна розмітка. В іноземній термінології вживається термін *part-of-speech tagging* (POS-tagging), дослівно – частиномовна розмітка.**
- **[S[N Nemo_NP1 ,_, [N the_AT killer_NN1 whale_NN1 N] ,_, [Fr[N who_PNQS N][V 'd_VHD grown_VVN [J too_RG big_JJ [P for_IF [N his_APP$ pool_NN1 [P on_II [N Clacton_NP1 Pier_NNL1 N]P]N]P]J]V]Fr]N] ,_, [V has_VHZ arrived_VVN safely_RR [P at_II [N his_APP$ new_JJ home_NN1 [P in_II [N Windsor_NP1 [safari_NN1 park_NNL1]N]P]N]P]V] ._. S]**
- **[http://ucrel.lancs.ac.uk/annotation.html]**

- **Синтаксична розмітка, що є результатом синтаксичного аналізу, або *парсинга* (*parsing*), виконуваного на основі даних морфологічного аналізу.**

- **Семантична розмітка. Хоча для семантики немає єдиної семантичної теорії, найчастіше семантичні теги позначають семантичні категорії, до яких відноситься дане слово або словосполучення, і більш вузькі підкатегорії, що специфікують його значення**

- **Анафорична розмітка. Фіксує референтні зв'язки, наприклад, займенникові.**

- **Просодична розмітка. У просодичних корпусах застосовуються мітки, що описують наголос та інтонацію. У корпусах усного розмовного мовлення просодична розмітка часто супроводжується так званою *дискурсною* розміткою, яка служить для позначення пауз, повторів, застережень, і т.д.**

Вимоги до розмітки

- **Розмітка повинна відповідати низці вимог, семи максимам Дж. Ліча [Leech 1997: 6-7].**

- **Розмітка мусить бути незалежною від тексту: повинна бути можливість прибрати розмітку і переглянути текст без неї, і, навпаки, вичленувати саму лише розмітку.**
- **Принципи розмітки, їх розробники та спосіб внесення розмітки в корпус повинні бути відомими кінцевому користувачу.**
- **Користувач повинен бути поставлений до відома про те, що розмітка не є безпомилковою, а являє собою лише потенційно корисний інструмент.**
- **В основу розмітки повинні бути покладені загальноприйняті і, по можливості, теоретично нейтральні лінгвістичні принципи.**

Реалізація будь-якого типу анотування передбачає низку процедур:

1. **Сегментизація тексту.**
2. **Формалізація параметрів анотування.**
3. **Створення тегсету чи набору формальних кодів з відповідною семантикою.**
4. **Визначення анотаційної схеми та її принципів.**

- **Автори монографії «Корпусна лінгвістика» [Корпусна лінгвістика , 2005: 51-53] зазначають такі критерії застосування стандарту:**

- **1) *Достатність*: набір структурних елементів повинен бути достатньо широким, щоб забезпечити хоча б більшість вимог. Водночас бажано, щоб схема розмітки не містила надлишкову інформацію.**
- **2) *Несуперечливість*: схема розмітки має бути сформована на базі несуперечливих правил, які б дозволяли однозначно визначити, які об’єкти належать до тегів, які – до атрибутів, що є вмістом тега тощо.**

- **3) *Відтворюваність*: схема кодування повинна ґрунтуватися на чітко визначених правилах, що дає можливість відтворити вихідний текст за допомогою простих алгоритмів.**
- **4) *Коректність*: за допомогою спеціального програмного забезпечення відбувається перевірка відповідності міток у документах їх структурним специфікаціям.**

- **5) *Можливість збору даних*: збір даних включає безпосереднє накопичення даних (за допомогою ручного вводу або з використанням автоматичного розпізнання тексту) та проведенням кодування даних.**
- **6) *Технологічність*: урахування потреб, пов'язаних з автоматичною обробкою текстів (вибір тексту згідно зі встановленими критеріями, використання спеціальних механізмів, типу міжтекстових покажчиків, поєднання текстів або інших елементів корпусу) тощо.**

- **7) *Можливість масштабування*: важливо, щоб будь-яка створена схема мала можливість поповнюватися.**
- **8) *Компактність*: проведення розмітки може істотно вплинути на розмір файлу, від чого залежить швидкість обробки даних текстів. Серед можливих методів досягнення компактності називають мінімізацію тегу, наприклад, пропущення або скорочення кінцевого тегу, застосування специфічних кінцевих тегів елементів або відмова від останніх; використання**

- **9) *Зрозумілість*: коли виникає потреба у безпосередній роботі користувача з текстом без використання спеціального програмного супроводу, прозорість розмітки є досить важливою.**

ЛЕКЦІЯ

Технологія створення корпусів

- **1. Визначення джерел лінгвального матеріалу.**
- **2. Введення даних.**
- **3. Попереднє опрацювання тексту.**
- **4. Конвертування й графематичний аналіз.**
- **5. Розмітка тексту.**
- **6. Коректування результатів автоматичної розмітки**

- **7. Конвертування розмічених текстів у структуру спеціалізованої лінгвістичної інформаційно-пошукової системи (corpus manager), що забезпечує швидкий багатоаспектний пошук і статистичну обробку.**
- **8. Забезпечення доступу до корпусу.**

Під час створення корпусу використовується низка процедур і програм, як-от: токенізація, лематизація, стеммінг, парсинг [Захаров 2011: 38-41].

Токенізація – це розбиття потоку символів природної мови на окремі значимі одиниці (токени, словоформи).

Лематизація – процес утворення початкової форми слова, виходячи з інших його словоформ. У багатьох мовах слово може зустрічатися в декількох формах з різними флексіями.

- **Стеммінг полягає в знаходженні стеми (основи) слова.**
- **Парсинг – це процес аналізу синтаксичної структури тексту чи частини тексту, що ґрунтується на зіставленні лінійної послідовності лексем (слів, токенів) мови з її формальною граматикою.**

Формати даних і стандартизація даних корпусу

- У цей час на основі міжнародного досвіду виробилися де-факто стандарти представлення метаданих, що базуються на описах текстів у рамках проекту Text Encoding Initiative (TEI) і на рекомендаціях EAGLES (Expert Advisory Group on Language Engineering Standards).

- **Стандарт TEI забезпечує оптимальну збалансованість між загальною моделлю подання природної мови і нескладною реалізацією кодування. Також TEI оперує великим набором засобів для подання як лінгвальної, так і металінгвальної інформації.**

- **У якості формальної мови розмітки широко застосовуються мови SGML (Standard Generelised Markup Language) і XML (Extensible Markup Language). У цей час стандарти EAGLES безпосередньо включаються в технологічне середовище мови XML, див., зокрема, розробку стандарту Corpus Encoding Standard for XML (XCES).**

Можливості використання корпусів у лінгвістичних дослідженнях

- ***Сфери застосування лінгвістичних корпусів***
- **Лексикографічні та граматичні дослідження на матеріалі корпусу**
- **Використання корпусів у навчанні іноземної мови**

- **(data-driven learning)**
- ***учнівські корпуси***

Дякую за увагу!!!!!!!!!!

Источник шаблона

ANIMATIONFACTORY

www.animationfactory.com

500 000 шаблонов PowerPoint, анимированных картинок, фоновых изображений и видеороликов для загрузки