УДК 81:33=161.2

С. Фокін, канд. філол. наук, доц.
Київський національний університет імені Тараса Шевченка, Київ

КОРПУСИ ТЕКСТІВ: ЗДОБУТКИ УКРАЇНИ ТА ПЕРСПЕКТИВИ ВРАХУВАННЯ ЗАКОРДОННОГО ДОСВІДУ

Розглянуто дев'ять текстових корпусів української мови, порівнюються їхні характеристики, можливості використання в дослідницькій роботі. З'ясовано, що найсуттєвішими параметрами електронних корпусів є розмітка як текстів у цілому (жанрово-тематична, ареальна, хронологічна, соціологічна), так і графічних слів у ньому (частиномовна, семантична); зараз бракує розмітки за дискурсивними характеристиками. Узагальнено принципи пошуку: можливість шукати слово, лексему, словосполучення, речення, а також маски виразів в узагальненому вигляді, однак виклик найближчого майбутнього – семантична розмітка та створення корпусів різних дискурсів.

Ключові слова: корпусна лінгвістика, корпус текстів, українська мова, метамовний пошук, методологія досліджень, дискурс.

Високі вимоги сучасності ставлять перед науковцями нові завдання. Автоматизація й розвиток штучного інтелекту полегшують рутинну роботу вчених, однак водночас вимагають дедалі серйозніших аналітичних зусиль в інтерпретації результатів після опрацювання масштабного фактичного матеріалу та вдосконалення і збільшення автоматизованої частини процесу, і філологічні науки не є винятком у цих процесах. Якщо раніше деякі складнощі обробки великих масивів текстового матеріалу видавалися нездоланними, натепер інформатизація, переведення в цифровий формат масштабних ресурсів зобов'язують лінгвістів, перекладознавців, літературознавців не відкладати розв'язання багатьох проблем, а шукати шляхи їхнього оптимального вирішення. Для цього необхідно доволі чітко усвідомлювати, якими можливостями вченого-філолога озброюють нові знаряддя, які цілі цілком досяжні вже у сучасному дослідженні, а які можливо досягти хіба що в перспективі. У дослідницькій роботі філологи послуговуються різними знаряддями автоматичної обробки тексту, однак найпоширеніші з них – програми опрацювання корпусів текстів, переважно конкордансери.

Незважаючи на десятилітню історію, використання корпусу української мови подекуди все ще сприймається як екзотичний і цілком новітній підхід. Виходячи зі сказаного, мета статті – охарактеризувати основні параметри текстових корпусних ресурсів української мови, їхні характеристики та потенціал у виконанні дослідницьких завдань, а також окреслити перспективи розвитку, ураховуючи світовий досвід.

Одним із перших корпусів на матеріалі українськомовних текстів став ***Корпус української мови***, розроблений ***лабораторією комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка*** [Корпус текстів української мови], у відкритому доступі з 2010 р.; у розробці брали участь співробітники лабораторії: Н. П. Дарчук (керівник проекту), В. М. Сорокін (програміст), О. Б. Сірук, Я. В. Ходаківська, Н. Г. Чейлитко, М. О. Лангенбах. Натепер якісно-кількісний склад корпусу такий: законодавчі тексти – 1 581 090 слововживань; наукові тексти – 8 712 314 слововживань; поетична мова – 799 408 слововживань; публіцистика – 40 681 393 слововживань; фольклорні тексти – 86 434 слововживань; художня проза – 53 953 337 слововживань. Це означає, що дослідник може обирати підкорпус за жанрово-стилістичними критеріями. Програмне забезпечення корпусу включає лематизатор, отже, можна проводити пошук не лише певної словоформи, а й лексеми в усій її парадигмі. У пошуковому діалоговому вікні можна вводити як одне слово, так і сполучення з двох слів.

Корпус текстів української мови дає змогу проводити пошук словосполучень (до двох слів) не лише формально, а й також частиномовної маски словосполучення. Скажімо, можна вибрати всі сполучення числівника з іменником або ж певного числівника з усіма іменниками. Докладніше про специфіку укладання корпусу, а також його потенціал для наукових розвідок можна дізнатися у статті Н. П. Дарчук "Дослідницький корпус української мови: основні засади і перспективи" [4]. Натепер Н. П. Дарчук працює над семантичною розміткою корпусу [5]. Семантична розмітка є серйозним викликом для сучасних лінгвістів з огляду на те, що семантичних груп у рази більше, ніж частин мови, і не існує їхньої однозначної класифікації.

На окрему згадку й увагу заслуговує ***Генеральний регіонально анотований корпус української мови*** [ГРАК], розроблений командою лінгвістів і програмістів з Києва, Осло (Норвегія), Єни (Німеччина): М. Шведова, Р. фон Вальденфельс, С. Яригін, М. Крук, А. Рисін, М. Возняк. Корпус охоплює тексти періоду з 1818 р. по 2018 р. і містить понад 25 тисяч текстів близько 3 850 авторів, а також тексти перекладів із 38 мов. Пошук слова чи виразу, побудова конкордансу в ньому можлива за такими критеріями: автор тексту, регіон, жанр тексту, місце публікації, роки публікації, мова оригіналу документа й ін. У корпусі можна здійснювати звичайний формальний пошук, пошук з "байдужими символами" (англ. – "wildcard characters"), а також запити *CQL ("Context Query Language",* спеціальна проста мова запитів, розроблена саме для пошуку в розміченому текстовому корпусі).

Корпусні знаряддя з часом набувають такої популярності, що в мережу потрапляють й аматорські програми, як, наприклад ***Корпус української мови*** або, як його характеризують самі упорядники, "Загальномовний (або національний) неанотований та несистематизований корпус української мови" [8]. Це знаряддя фактично працює як пошуковий рядок сайту електронної бібліотеки. Корпус цілком придатний для проведення первинного пошуку, що не вимагає точних підрахунків.

Багатьох дослідників має зацікавити ***Архів української періодики онлайн***. Як зазначено на сторінці ресурсу, "наразі оцифровано близько 700 000 сторінок понад 400 видань українською, польською, німецькою, румунською, їдиш, кримсько-татарською та російською мовами, що видавалися у різних регіонах України та поза її межами від початку до 50-х років XX ст. Ресурс постійно наповнюється" [1].

Чимало сучасних сайтів містить текстові ресурси, об'єднані за певним тематичним, функціональним чи іншим критерієм: законодавчі тексти на сайтах парламентів, тексти статей журналів і газет, бібліотеки художньої, наукової студентської літератури, зібрання творів певного автора. Кожен із таких ресурсів можна використовувати як простий корпус. Щоправда, у таких текстах можливо здійснювати лише простий форма-

льний пошук сполучення літер або слів, а обирати додаткові параметри та/або використовувати маску пошуку неможливо. Тому ***спеціальний анотований корпус, обмежений творами певного автора***, навіть менший за обсягом, становитиме для дослідника неабияку цінність. В Україні цією проблематикою глибоко займається дослідниця з університету "Львівська політехніка" С. Бук, і в статті робить ретельний огляд різних корпусних ресурсів, зокрема й тих, що укладені вручну на матеріалі творів українських авторів: "До недавнього часу в українській лінгвістиці домінувала ручна методика укладання словників, зокрема, у письменницькій лексикографії. Цим способом було створено словники мови Т. Шевченка та Г. Квітки-Основ'яненка; словопокажчики творів І. Котляревського, Т. Шевченка, Л. Українки, А. Тесленка, В. Стефаника, Ю. Федьковича; глосарій поетичної мови Василя Стуса" [2, с. 73–74]. Серед параметрів, необхідних у пошуку слів у корпусі, С. Бук включає до ресурсу різні характеристики ідіостилю, у тому числі частотний словник автора як у цілому, так і в окремому творі, і в обраній низці творів, і в певному жанрі, а також фіксацію часового періоду вживання тієї чи іншої лексеми та інш. [2, с. 81–82]. Звичайно, такою детальною параметризацією дослідника не забезпечив би простий формальний пошук ані в текстах письменника, ані в загальних корпусах великого обсягу.

До деяких корпусів українськомовних текстів наразі немає публічного доступу, і з огляду на їхній потенціал для дослідників у найближчому майбутньому назвемо їх теж. ***Український національний лінгвістичний корпус***, нині доступний тільки зареєстрованим користувачам, потребує інсталяції на персональний комп'ютер, розроблений Українським мовно-інформаційним фондом НАН України. ***Корпус текстів української мови.*** Раніше був доступний, як документує Н. П. Дарчук, Корпус текстів української мови (за адресою http://corpora.pp.ua, розробник – кафедра української мови і прикладної лінгвістики Донецького національного університету [6]). Автор статті (у співавторстві з М. С. Івановим) працює над розробкою ***конкордансера для опрацювання корпусу співвідносних текстів "TexPeer",*** основне призначення якого – проведення контрастивних досліджень; наразі корпус наповнено співвідносними за темами текстами іспанської та української мов обсягом по 500 000 слововживань із кожної мови.

Корпуси української мови містяться також у складі колекцій корпусів. Серед найвідоміших колекцій корпусів, що одними з першими з'являються в пошуковому рядку, слід назвати корпуси Лейпцизького університету [Corpora Collection of Leipzig University] і корпуси Лідського університету [Leeds corpora collection]. Зазначені ресурси характеризуються вражаючо масштабним обсягом. Зокрема, у корпусах Лейпцизького корпусу українськомовна частина становить 102 429 857 речень, отже, загальна кількість словоформ може сягати мільярда.

Цінність корпусу більшою мірою обумовлена його якісними, ніж кількісними параметрами. Так, Г. Енґвел у статті "Не випадковість, а свідомий вибір: критерії укладання корпусу" ("*Not chance but choice: criteria in corpus creation")*, розмежовує поняття "загального корпусу" та "якісного корпусу": "У корпусній лінгвістиці те, що більше, не обов'язково краще. Дрібний, однак добре організований корпус може виявитись кориснішим, ніж більший за обсягом, однак менш послідовний за змістом (...). За своїм досвідом я переконався, що корпуси від 20 000 до 200 000 слововживань виявились цілком придатними для здійснення дослідження. Більші корпуси укладаються надто довго, а менші містять недостатньо даних для належної інтерпретації" [11, с. 352]. Звичайно, для дослідження більшості лексичних і граматичних явищ зазначених обсягів може виявитися достатньо; такі явища, як неологізми, окказіоналізми, мовні аномалії можуть зустрічатися відносно рідко, і малий корпус навряд чи дасть повну картину їхніх властивостей. Однак корпуси малого обсягу мають свої переваги, які докладніше розглядаємо в статті [10].

Кількість корпусів зростає щороку, тому неможливо навіть приблизно осягнути всі можливі ресурси. Однак лише в незначній частині сучасних корпусів містяться соціолінгвістичні дані про авторів (наявність/відсутність білінгвізму, стать, освіта). На наш погляд, різнобій у форматах і наповненнях корпусів можна пояснити тим фактом, що переважна більшість запитів, досліджень, що ґрунтуються на певному корпусі, найчастіше спрямовані не на якусь мову загалом, а на певний дискурс, що виокремлюється за тематичними, авторськими, прагматичними чи іншими позамовними чинниками; позаяк кількість дискурсів теоретично не обмежена, відповідно, і кількість корпусів певної мови може суттєво розростатися. Можна прогнозувати, що в майбутньому процес укладання корпусів суттєво зміниться: замість пошуку й підбирання текстів переглядатимуться вже наявні корпуси та збагачуватися додатковими, цінними для дослідження, параметрами вже наявні тексти, що такої розмітки раніше не мали. У такий спосіб можна буде легко виокремлювати спеціальний підкорпус із загальних корпусів. Сучасні програми опрацювання текстів здатні автоматично визначати тему тексту (напр., розробки Н. П. Дарчук), однак викликом для корпусної лінгвістики на завтра є автоматизоване визначення типу дискурсу, до якого належить текст. Більше того, необхідно порушувати питання взагалі про досяжність цієї мети, оскільки дискурси виникають довкола тієї чи іншої сфери життя, як, наприклад, "дискурс катастроф", "ядерний дискурс", "корпоративний дискурс" тощо. Будь там як, саме поняття дискурсу, натепер дуже розлоге, об'ємне й неоднозначне, теж потребує уточнення в межах корпусної лінгвістики.

Отже, потрібно вже зараз розробляти можливу методику розмітки й аналізувати ті параметри та характеристики, що цікавлять (реально й потенційно) дослідників. Дуже детально розмітку різного роду метаданих розглядає Л. Бурнард [12]. Зокрема, лінгвістична розмітка за різними мовними рівнями й категоріями (аналітичні метадані), описові дані, із-поміж яких у тому числі й демографічні характеристики кожного мовця. Можна очікувати на появу найближчим часом відповідного стандартну *ISO*. Наявність стандарту не обов'язково забезпечить уніфікацію. Адже в окремих розвідках досить буде обмежитися достатньо простими корпусами з мінімальною розміткою. До того ж завжди залишається на порядку денному всіх наук питання неоднозначності класифікацій: це стосується всіх явищ, які людина здатна аналітично розподіляти на групи: від елементарних часток, спектрів кольорів електромагнітного випромінювання до стилів, жанрів текстів і дискурсів. Якщо доволі тривалий час була поширеною класифікація текстів на п'ять функціональних стилів, за В. В. Виноградовим, нині додаються альтернативні класифікації, що включають і військовий, і конфесійний стилі. Не вдаючись до аргументів на користь чи проти, підкреслимо, що зміна парадигми в теорії передбачає суттєву переробку баз даних на практиці. Більше того, може виявитися, що один і той самий об'єкт, явище, характеристика може підпадати під декілька класифікацій одночасно, так само й певний текст за тематичним критерієм можна від-

нести і до філософських, і медичних, і релігійних водночас, не говорячи вже про те, що один і той самий текст може одночасно належати до туристичного, медичного, законодавчого й корпоративного дискурсів. Інформатики можуть частково вирішувати подібні казуси класифікації за допомогою таких знарядь, як "muli-label classificaction". Не вдаючись у деталі, прокоментуємо, що багатокатегорійна класифікація об'єктів при всіх своїх перевагах може на декілька порядків уповільнювати реалізацію алгоритму або навіть його практично унеможливлювати. Аналітичність штучно проводить межі між виявами певних явищ, і будь-яка зміна підходу обумовить необхідність переробки, виправлення або створення нової емпіричної бази. Тому можна прогнозувати, що корпуси текстів будуть постійно створюватися, оновлюватися, реорганізовуватися, доповнюватися, і не передбачається того моменту, коли можна буде з полегшенням поставити крапку в цій справі.

Окрім власне текстового наповнення відповідно до тих чи інших кількісних і якісних параметрів, інтерфейс програм-конкордансерів, що забезпечують доступ до текстів і їхніх фрагментів, теж суттєво варіюється. Майже всі конкордансери показують абсолютну частоту вживання словоформ, що дає змогу укладати частотні словники словоформ, а, за наявності лематизатора – і частотні словники лексем; відносну та середню частоту програми розраховують рідше. У намаганні заповнити цю прогалину розробники програми-конкордансера "TexPeer″ включили до неї функцію розрахунку стандартного відхилення для перевірки достовірності вибірки.

Проривом у сфері інформатизації української мови можна вважати проект "lang-uk", що ним керує Д. Чаплинський. На сайті проекту містяться текстові корпусні ресурси українських періодичних видань, законодавчих текстів, а також український браунівський корпус [9], а також інші корисні ресурси, однак користувачеві доведеться подбати про встановлення або написання відповідного програмного забезпечення корпусів. Зазначимо, що "браунівським" у корпусній лінгвістиці називається корпус, укладений за критеріями першого текстового корпусу у Браунівському університеті (США), що містив 500 000 текстів по 2 000 слововживань кожний. Український "браунівський" корпус зараз у розробці, яка відкрита на сайті проекту "lang-uk" для всіх, хто бажає долучитися.

Підсумовуючи якісні та кількісні параметри корпусів текстів української мови, варто зазначити, що потенційно значущі характеристики – тема, час, жанр, прагмастилістичні параметри, соціолінгвістичні дані автора, обсяг та інші – обумовлюють структуру, обсяг і призначення корпусу. Корпусні текстові ресурси можна розподілити за такими критеріями: загальні корпуси та спеціалізовані (на певному матеріалі, темі, авторі). Кожну з цих категорій варто також розділити на корпуси без розмітки та корпуси з розміткою (за жанрово-стилістичними, тематичними, ареальними, соціолокальними й іншими критеріями текстів, частиномовною і семантичною розміткою словоформ, що нині є завданням лише частково вирішеним). З огляду на елементарну автоматизацію простого формального пошуку в цифрових текстових документах, загальним некласифікованим корпусом може стати будь-який сайт із текстами певної мови тематичного або іншого спрямування, як бібліотека української літератури, архів української періодики.

Наявність декількох корпусів української мови ще не означає вичерпного охоплення зрізу мови в усіх її можливих вимірах: дослідникам потрібні не лише стандартні і правильні тексти, а й тексти неносіїв мови з типовими й нетиповими помилками (до прикладу, корпус *Corpus de Aprendientes de Español* (корпус неносіїв іспанської мови) [13]), тексти перекладів певною мовою, окремо необхідні усні й письмові тексти; спеціальним предметом дослідження стають інтернетні тексти, які теж можуть бути й усними і письмовими, бездоганними з погляду грамотності або недосконалими; корпуси певної спрямованості теж стануть у пригоді дослідникам: окрім регіонального корпусу, на часі питання про створення акцентологічного корпусу, звукового корпусу, корпусу текстів кінофільмів, протокольних промов, у межах яких теж можлива й доцільна детальна класифікація текстів на типи й підтипи. Дискурсивне спрямування корпусів текстів теж необхідне, так само як і уточнення поняття "дискурс" у межах корпусної лінгвістики, а також вирішення питання щодо практичної можливості автоматизованого визначення типу дискурсу. Значення для дослідження має не лише обсяг корпусу (для багатьох розвідок цілком достатньо корпусу порядку сотні тисяч слововживань), а навіть більшою мірою якісна розмітка текстових документів і його складових.

Список використаних джерел

1. Архів української періодики онлайн. URL : http://uk.glosbe.com/ https://libraria.ua/

2. Бук С. Корпус текстів Івана Франка: спроба визначення основних параметрів / С .Бук // Прикладна лінгвістика та лінгвістичні технології: MegaLing 2006 : зб. наук. пр. – К. : Довіра, 2007. – С. 72–82.

3. ГРАК, Генеральний регіонально анотований корпус української мови / М. Шведова, Р. Фон Вальденфельс, С. Яригін, М. Крук, А. Рисін, М. Возняк. – К. ; Осло ; Єна, 2017–2018. – URL : uacorpus.org.

4. Дарчук Н. П. Дослідницький корпус української мови : основні засади і перспективи / Н. П. Дарчук // Вісн. Київ. нац. ун-ту ім. Т. Шевченка. Серія: Літературознавство. Мовознавство. Фольклористика. – К. : ВПЦ "Київський університет", 2010. – № 21. – С. 45–49.

5. Дарчук Н. П. Можливості семантичної розмітки корпусу української мови (КУМ) / Н. П. Дарчук // Науковий часопис Нац. пед. ун-ту імені М. П. Драгоманова. Серія 9 : Сучасні тенденції розвитку мов: зб. наук. праць. – К. : Вид-во НПУ імені М. П. Драгоманова, 2017. – Вип. 15. – С. 18–28.

6. Дарчук Н. П. Паралельний корпус текстів ПарКУМ. Драгоманова / Н. П. Дарчук, М. О. Лангенбах, В. М. Сорокін, Я. В. Ходаківська . Серія 9. Сучасні тенденції розвитку мов. – К. : Вид-во НПУ імені М. П. Драгоманова, 2017. – Вип. 15. – С. 28–35.

7. Корпус текстів української мови. – URL : http://www.mova.info/corpus.aspx?l1=209

8. Корпус української мови. – URL : http://korpus.org.ua/

9. Проект lang-uk. – URL : http://lang.org.ua/uk/corpora/

10. Фокін С. Б. Переваги корпусів малого обсягу для досліджень макроявищ у перекладі / С. Б. Фокін // Мовні і концептуальні картини світу. – К. : ВПЦ "Київський університет", 2015. – Вип. 51. – С. 590–597.

11. Bowker L. Towards a Methodology for a Corpus Based Approach to Translation Evaluation / L. Bowker // Meta. – 2001. – Vol. 46 (2). – № 2.

12. Burnard L. Metadata for Corpus Work / L. Burnard. – URL : http://users.ox.ac.uk/~lou/wip/metadata.html#HDR

13. CAES, Corpus de Aprendices de Español. – URL : http://galvan.usc.es/caes/search.

14. Corpora Collection of Leipzig University. – URL : http://corpora.uni-leipzig.de.

15. Leeds Collection of Internet Corpora. – URL : http://corpus.leeds.ac.uk/internet.html.

References

1. Arkhiv ukrayinskoyi periodyky onlajn. URL : http://uk.glosbe.com/ https://libraria.ua/

2. Buk S. Korpus tekstiv Ivana Franka: sproba vyznachennya osnovnykh parametriv // Prykladna linhvistyka ta linhvistychni tekhnolohiyi: MegaLing 2006: zb. nauk. pr.– K. : Dovira, 2007.– S. 72–82.

3. GRAK, General Regionally Annotated Corpus of Ukrainian / M. Shvedova, R. Fon Valdenfels, S. Yaryhin, M. Kruk, A. Rysin, M. Voznyak. – Kyyiv, Oslo, Jena, 2017-2018. URL : uacorpus.org

4. Darchuk N. P. Doslidnyckyj korpus ukrayinskoyi movy: osnovni zasady i perspektyvy // Visnyk Kyyivskoho nac. un-tu im. T. Shevchenka. Seriya: Literaturoznavstvo. Movoznavstvo. Folklorystyka. – K.: VPC “Kyyivskyj universytet”, 2010. No 21. S. 45–49.

5. Darchuk N. P. Mozhlyvosti semantychnoyi rozmitky korpusu ukrayinskoyi movy (KUM) // Naukovyj chasopys Nacionalnoho pedahohichnoho universytetu imeni M.P. Drahomanova. Seriya 9 : Suchasni tendenciyi rozvytku mov: zb. nauk. prac. – K. : Vyd-vo NPU imeni M. P. Drahomanova, 2017. Vyp. 15. S. 18–28.

6. Darchuk N. P., Lanhenbakh M. O., Sorokin V. M., Khodakivska Ya. V. Paralelnyj korpus tekstiv ParKUM. Drahomanova. Seriya 9. Suchasni tendenciyi rozvytku mov. 2017. Vyp. 15. S. 28–35.

7. Korpus tekstiv ukrayinskoyi movy. URL : http://www.mova.info/corpus.aspx?l1=209
8. Korpus ukrayinskoyi movy. URL : http://korpus.org.ua/
9. Proekt lang-uk. URL : http://lang.org.ua/uk/corpora/
10. Fokin S. B. Perevahy korpusiv maloho obsyahu dlya doslidzhen makroyavyshh u perekladi // Movni i konceptualni kartyny svitu. K. : VPC "Kyyivskyj universytet", 2015. Vyp. 51. S. 590–597.
11. Bowker L. Towards a Methodology for a Corpus Based Approach to Translation Evaluation / L. Bowker // Meta, 2001, 46, #2.
12. Burnard L. Metadata for Corpus Work / L. Burnard. URL : http://users.ox.ac.uk/~lou/wip/metadata.html#HDR
13. CAES, Corpus de Aprendices de Español. URL : http://galvan.usc.es/caes/search
14. Corpora Collection of Leipzig University. URL : http://corpora.uni-leipzig.de
15. Leeds Collection of Internet Corpora. URL : http://corpus.leeds.ac.uk/internet.html

Надійшла до редколегії 11.10.18

S. Fokin, PhD, Associate Professor
Taras Shevchenko National University of Kyiv, Kyiv

TEXTUAL CORPORA: UKRAINIAN LINGUISTS' ACHIEVEMENTS AND ASSIMILATION OF FOREIGN EXPERIENCE

Though five or more corpora of the Ukrainian language exist since 2010 or earlier, the majority of them remain unknown to researchers and corpus-based studies in Ukrainian philology are seen rather as exotic and exceptional cases. In the present study we offer an overview of nine Ukrainian corpora, among which the widest and the fullest are "Ukrainian Language Corpus" at web-portal mova.info and "GRAK" ("General Regionally Annotated Corpus of Ukrainian). Two of them make part of corpora collections ("Leeds Corpora Collection") and ("Corpora Collection of Leipzig University"); two corpora are made on the basis of electronic document archives, which appears to demonstrate that nowadays any set of electronic textual documents corresponding to a common criterion are convertible into a simple corpus. Today's large corpora provide the possibilities of making searching queries according to multiple criteria: subject, period, style, gender, sociolect, etc. Another useful feature of large general corpora is searching by means of regular expressions and "wildcard symbols" which provide the possibility of making a set of queries at once corresponding to a certain search mask. These formats vary from one corpus to another: from asterisks to classical regular expressions and CQL queries; not all Ukrainian corpora still offer the possibility of searching phrases larger than two words. Most large corpora are POS-annotated, which is a great achievement of computational linguistic, but the actual challenge is the semantic annotation, which is being developed at present for the corpus at mova.info. Among the parameters of the corpus those which matter for a research are, besides its volume (many case studies do not require a corpus larger than a hundred thousand tokens), the document annotation and text components annotation. As more and more researches are discourse-oriented, it is to anticipate that more discourse-oriented corpora will be appearing in the near future.

Keywords: corpus linguistics, textual corpus, Ukrainian language, metalinguistic search, researching methodology, discourse.