**Література:**
1. Anderson Vasby K. Governing Codes : Gender, Metaphor, and Political Identity / K. Anderson Vasby, K. Horn Sheeler. – Lexington Books, 2005. – 243 c.
2. Bock von Wülfingen B. Genetisierung der Zeugung : eine Diskurs- und Metaphernanalyse reproduktionsgenetischer Zukünfte / B. Bock von Wülfingen., 2007. – 370 c.
3. Ebeling S. Geschlechterforschung und Naturwissenschaften. Einführung in ein komplexes Wechselspiel / S. Ebeling, S. Schmitz. – Wiesbaden : Verlag, 2006. – 392 c.
4. Koller V. Analyzing metaphor and gender in discourse / Koller // Unité et diversité de la linguistique / V. Koller – Lyon : Atelier intégré de publication de l'Université Jean Moulin – Lyon 3, 2011. – С. 125–158.
5. Schmitt R. Metaphernanalyse und die Konstruktion von Geschlecht / R. Schmitt // Forum Qualitative Sozialforschung. – 2009. – № 10.

*УДК 81`33*

***О. Ю. Плахотнікова,***
*Київський національний університет імені Тараса Шевченка, м. Київ*

## СУЧАСНИЙ СТАН КОРПУСНИХ ДОСЛІДЖЕНЬ В УКРАЇНІ

*У статті проаналізовано основні напрямки сучасних корпусних досліджень в Україні та охарактеризовано найвідоміші створені одномовні корпуси писемного й усного мовлення. Також описано проект Корпусу усного транскрибованого українського мовлення.*

***Ключові слова:*** *корпусна лінгвістика, корпус писемного мовлення, корпус усного мовлення, анотація.*

***СОВРЕМЕННОЕ СОСТОЯНИЕ КОРПУСНЫХ ИССЛЕДОВАНИЙ В УКРАИНЕ***

*В статье проанализированы основные направления современных корпусных исследований в Украине и охарактеризованы наиболее известные существующие одноязычные корпуса письменной и устной речи. Также описан проект Корпуса устной транскрибированной украинской речи.*

***Ключевые слова:*** *корпусная лингвистика, корпус письменной речи, корпус устной речи, аннотация.*

***MODERN STATE OF CORPUS RESEARCH IN UKRAINE***

*The theoretical problems of corpus linguistics have been in the focus of attention of many Ukrainian researchers, but few papers reported on the contemporary state of corpus studies in Ukraine. The article analyses the main trends of research within the field of modern corpus linguistics in Ukraine. It characterizes the most famous existing monolingual corpora of written and spoken Ukrainian language. Such summary of the state of modern corpus resources has to contribute to create and update many corpus databases. Nowadays in Ukraine there is an increasing need for linguistic speech corpora containing phonetic markup. The author also describes the main principles of the transcribed Ukrainian speech corpus. The transcribed Ukrainian speech corpus is a scientific project on the base of the Experimental Phonetics Educational Laboratory, Institute of Philology, National Taras Shevchenko University of Kyiv. Consequently, according to the description of corpus research in Ukraine, we can make a conclusion that nowadays there are several written and spoken Ukrainian language corpora which are of the big scientific interest, and can be a good material for scientific research and different linguistic studies. Nevertheless, it is very important to create Ukrainian speech corpora that give the opportunity to analyse the sound implementation of the recorded texts.*

***Keywords:*** *corpus linguistics, written language corpus, spoken language corpus, annotation.*

Корпусний метод і його ресурси сьогодні активно використовують у різних лінгвістичних дослідженнях багатьох світових мов; мовні корпуси розрізняють за обсягом, типом, структурою, наповненням, призначенням тощо. Масиви даних писемного й усного мовлення дають змогу як досліджувати окремі мовні явища, так і з'ясовувати закономірності функціонування мовних одиниць різних рівнів. На думку О. Демської (Кульчицької), у сучасному українському мовознавстві вже присутні основні теоретичні й практичні передумови розвитку корпусної лінгвістики.

Теоретичні проблеми корпусної лінгвістики перебувають у центрі уваги українських учених О. Демської (Кульчицької), В. Широкова, Є. Карпіловської, Н. Дарчук, В. Жуковської, В. Балог, С. Бук та ін. У 2005 році побачили світ колективна монографія «Корпусна лінгвістика» (автори – В. Широков, О. Бугаков, Т. Грязнухіна та ін.) і монографія О. Демської-Кульчицької «Основи Національного корпусу української мови», що стали теоретичною базою для наступних корпусних досліджень. В Україні корпусними студіями активно займаються вчені Інституту української мови НАНУ та Інституту мовознавства ім. О. Потебні НАНУ, Українського мовно-інформаційного фонду НАНУ, Інституту філології Київського національного університету імені Тараса Шевченка, лабораторії комп'ютерної лінгвістики Київського національного лінгвістичного університету, Львівського національного університету імені Івана Франка тощо.

Метою статті є аналіз основних досягнень українських учених у галузі корпусної лінгвістики та огляд найвідоміших створених одномовних корпусів писемного й усного мовлення. Такий зріз сучасного стану корпусних ресурсів має сприяти побудові й оновленню корпусних баз даних, зокрема й проекту Корпусу усного транскрибованого українського мовлення (КУТУМ). Це зумовлює актуальність нашої роботи.

Перш за все відзначимо, що українська наука потребує загальномовних корпусів української мови. Один з таких проектів утілює в життя Український мовно-інформаційний фонд НАНУ. На основі Національної словникової бази Фонд під керівництвом академіка НАН України В. Широкова створив репрезентативний Український національний лінгвістичний корпус (УНЛК), що проектувався як фундаментальна лексикографічна мовно-інформаційна система підтримки мовознавчих досліджень з орієнтацією на створення лексикографічних продуктів. Об'єктом дослідження в УНЛК є сучасна українська літературна мова. Ця система нараховує понад 110 млн. слововживань, які містяться в джерелах електронних українських текстів художньої, наукової, конфесійної літератури, публіцистики, ділових, законодавчих документів XIX-XXI століть. Для джерельної бази даних Корпусу впроваджено Формат бібліографічного опису. Починаючи з 2000 р., було проведено роботу зі створення мережевої версії цифрової бібліотеки Українського мовно-інформаційного фонду, яка є складовою частиною УНЛК [11]. УНЛК – це ресурс закритого типу, доступ до якого надається в межах наукової співпраці Фонду з освітніми установами.

На початку XXI ст. відділ лексикології та комп'ютерної лексикографії Інституту української мови НАН України розробляв ідею Національного корпусу української мови. У межах роботи над цим проектом було опубліковано монографію

О. Демської-Кульчицької «Основи Національного корпусу української мови», присвячену «проблемам теорії та практики створення електронного текстового корпусу загалом і корпусу української мови зокрема»; тут також вміщено розроблену «систему кодів для морфологічної анотації саме корпусних текстів сучасної української мови з урахуванням міжнародних корпусних анотаційних стандартів» [4, с. 9]. Через кілька років, після бурхливої наукової дискусії, доопрацювання та доповнення праці назву наукового проекту було змінено на «Корпус сучасної української мови». У результаті зміни загальної концепції в 2011 році побачила світ монографія «Текстовий корпус: ідея іншої форми» О. Демської (Кульчицької). Мета створення Корпусу сучасної української мови – це побудова емпіричної дослідницької моделі сучасної української мови. За визначенням О. Демської (Кульчицької), «предметну галузь дослідницького, загальномовного, фрагментного, динамічного, синхронного, одномовного Корпусу сучасної української мови становитиме сучасна українська мова у таких її формах: літературна мова, діалекти та частково соціолекти» [4, с. 196]. Нижньою кількісною межею обсягу Корпусу автор вважає один мільйон слововживань, а верхню, із застосуванням прийому пропорційного збільшення, планує на сто мільйонів.

Один із відомих електронних текстових ресурсів української мови наукового характеру й відкритого типу створений відділом структурно-математичної лінгвістики Інституту мовознавства ім. О. Потебні НАН України. Результатом багаторічних досліджень наукового колективу відділу стала повнотекстова база даних мовного фонду, що містить близько 700 тис. слововживань, забезпечена функціями орфографічного контролю текстів, аналізу їхньої семантичної, морфологічної та синтаксичної структури. Для роботи з цими масивами даних розроблено спеціальні текстові процесори, до складу яких увійшли системи автоматичного орфографічного контролю й редагування текстів РУТА і машинного українсько-російського та російсько-українського перекладу ПЛАЙ (автори – Т. Грязнухіна, Л. Орлова, Н. Дарчук, В. Критська та ін.) [4, с. 14; 5, с. 78]. Результатом роботи вказаного відділу було створення генерального реєстру комп'ютерного морфемно-словотвірного фонду української мови обсягом 171 304 слова з інформацією про їхню морфемну будову, частиномовну належність, кількість властивих їм значень, абсолютну частоту вживання в півмільйонній текстовій вибірці сучасної української художньої прози [5, с. 79]. Для роботи з указаним фондом розроблено спеціальну систему «Морфолог», а сам морфемно-словотвірний фонд можна вважати важливим етапом лінгвоукраїністики на шляху до корпусної лінгвістики у практичному й теоретичному планах [4, с 15].

У межах іншого наукового проекту «Корпусу української мови» колектив фахівців навчально-наукової лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка розробив Дослідницький корпус сучасної української мови, тестова електронна версія якого представлена на лінгвістичному порталі Mova.info [7]; її обсяг – майже 50 млн. словоформ, масив даних працює як інформаційно-довідкова система. При анотуванні корпусу було використано третій варіант системи автоматичного граматичного аналізу тексту (АГАТ) – новий контекстний аналіз, створений співробітниками вказаної лабораторії комп'ютерної лінгвістики Н. Дарчук і В. Сорокіним; базу даних анотовано на рівні морфеміки, морфології й синтаксису. Цей корпус дає змогу статистично обґрунтувати дослідження з української мови, літератури, культури, поглибити його відповідно до сучасних наукових норм [7]. На сайті Mova.info є можливість отримувати різноаспектну статистичну інформацію про мовну одиницю за допомогою частотних словників слів і словоформ.

Колектив кафедри української мови та прикладної лінгвістики Донецького національного університету створив Корпус текстів української мови із морфо- і метарозміткою з метою вивчення граматичної службовості (обсяг бази даних – близько 5 мільйонів слововживань). У межах проекту реалізовано зокрема систему тегів для службових частин мови [3].

Над проблемою створення та використання корпусів текстів працює науковий колектив лексикографічної лабораторії LEXILAB на факультеті романо-германських мов Національного університету «Острозька академія». Проект передбачає роботу над створенням корпусу української англійської мови (з текстами, представленими англійською мовою українцями), корпусу міжнародної української мови (це тексти, представлені українською мовою іноземцями, що не є громадянами України) та корпусу спеціалізованого наукового мовлення (у вигляді корпусу наукових статей) [6].

Дослідники С. Бук і А. Ровенчак із Львівського національного університету імені Івана Франка працюють над зовнішньо і внутрішньо анотованим Корпусом текстів Івана Франка (КТФ); проектом передбачено охопити всі твори І. Франка (приблизно 7 млн. слововживань) з усіма особливостями фонетичного, морфологічного, словотвірного, лексико-семантичного, фразеологічного, синтаксичного рівнів. КТФ має репрезентувати підсистему західного варіанту української мови кінця XIX – початку XX ст. у різних стилях мови [1, с. 63]. Уже створено корпус текстів роману І. Франка «Перехресні стежки», на основі якого укладено частотний словник і конкорданс.

На сьогодні в Україні вже створено чимало корпусів текстів писемного мовлення, але існує невелика кількість лінгвістичних корпусів усного мовлення, які містять фонетичну розмітку. Мовленнєві бази даних є здебільшого ресурсами закритого типу, що обслуговують потреби конкретних наукових установ. Отже, важливим завданням є створення саме корпусів усного українського мовлення, що дають можливість опрацьовувати звукову реалізацію записаних текстів.

У галузі розпізнавання та синтезу українського мовлення беззаперечним лідером в Україні є відділ розпізнавання та синтезу звукових образів Міжнародного науково-навчального центру інформаційних технологій та систем (МННЦІТС) у м. Києві й Українська асоціація з оброблення інформації та розпізнавання образів (УАсОІРО) при ньому [10]. Для робіт у галузі автоматичного розпізнавання українського мовлення дослідниками було розроблено такі мовленнєві корпуси:

1. **UkReco** – українськомовний багатодикторний мовленнєвий корпус, що містить понад 30 000 реалізацій слів і тисячі речень (залучено близько 100 дикторів із різних областей України). Реалізації слів зберігають частотні пропорції фонем і є фонетично збалансованими, при доборі слів враховували їхні частотні характеристики. Цей корпус використовують у дослідженнях з розпізнавання ізольованих слів, адаптації до голосу диктора та при побудові акустичних моделей для усного словника-перекладача [2, с. 56]. Мовленнєву базу даних **UkReco** було створено к.т.н. М. Сажком.

2. Інший корпус містить аудіотексти виступів **депутатів Верховної Ради України**, записані через телевізійну мережу від приблизно 400 дикторів. Особливості цього масиву даних: спонтанне мовлення, швидкий темп, емоційна забарвленість, висока якість запису. Обсяг – приблизно 40 годин [2, с. 56]. Ці записи використовуються для розробки експериментальної системи розпізнавання парламентського мовлення і вдосконалення авторської програми FindRealTranskription.

3. Постійно доповнюється і вдосконалюється **Акустичний корпус українського ефірного мовлення (АКУЕМ)**, що містить читане, підготоване та спонтанне теле- і радіомовлення (більша частина – це спонтанне мовлення), а також публічне мовлення і мовлення в природному середовищі. АКУЕМ містить понад 300 годин анотованого мовлення, записаного приблизно від 2000 дикторів. Словник цього корпусу налічує понад 65 000 слів української мови [9, с. 81–82].

Серед призначених для синтезу українського мовлення корпусів, розроблених МННЦІТС, привертає увагу база даних, створена на основі жіночого голосу професійного диктора. Також у відділі втілено в життя проект українськомовного

корпусу опорного диктора обсягом понад 50 годин мовлення, який використовується в дослідженнях пофонемного і поскладового розпізнавання [2, с. 56].

Що ж до корпусів усного мовлення, записаного в природних умовах, то разом з Українською асоціацією з оброблення інформації та розпізнавання образів Київський МННЦІТС працює над **телефонною мовленнєвою базою даних** спонтанного російського та українського мовлення обсягом близько 5 Гб (у GSM форматі), яка поки що не анотована. Вона містить реальні записи з різних мобільних телефонів, частота дискретизації яких – 8 000 Гц [10].

Працюють над створенням мовленнєвих баз даних і українські діалектологи. Варто відзначити корпус діалектного мовлення «Українські говірки Донеччини», який створено колективом співробітників кафедри української мови Донецького національного університету (Л. Фроляк, З. Омельченко, В. Познанська, Н. Клименко, Н. Михайлова, В. Дроботенко). Цей матеріал було впорядковано та видано в 2000 році на компакт-диску «Українські говірки Донеччини. Фонотека. Вип. 1» (це комп'ютерна хрестоматія-фонотека у звуковій та графічній формах). Для роботи з указаними корпусами аудіотекстів створено доступний і зручний для користувачів-філологів інтерфейс. Фонотека «Українські говірки Донеччини» містить понад 50 годин звучання діалектних текстів спонтанного мовлення, записаних від інформантів у 65 населених пунктах Донецької області в 1997-2000 роках. Аудіотексти збережено у форматі WAV.compress 24 KBit/sec. (24 000 Hz Mono), вони супроводжуються передмовою про історію формування українських говірок на Донеччині та їхні структурно-граматичні особливості [4, с. 16]. Такі записи допомагають виявити фонетичні особливості говірки (звукову систему, інтонацію, наголос тощо).

На базі навчальної лабораторії експериментальної фонетики Інституту філології Київського національного університету імені Тараса Шевченка ми створюємо власний Корпус усного транскрибованого українського мовлення з фонетичною розміткою, а теоретичною метою нашого дослідження є напрацювання теоретичних підходів і процедур для управління корпусними ресурсами українського усного мовлення.

Для створення фрагментів вказаного корпусу й анотування аудіофайлів використано комп'ютерну програму ELAN [13], її безкоштовно поширює Інститут психолінгвістики імені Макса Планка у місті Неймеген (Нідерланди) [12]. Застосування програми ELAN уможливлює створення багаторівневих анотацій медіафайлів Корпусу усного транскрибованого українського мовлення. Матеріалом нашого дослідження слугувало українське природне усне мовлення, а саме фрагменти звукового корпусу тривалістю понад 100 хвилин: це 27 аудіозаписів у **wav**-форматі трьох студентів, п'ятьох викладачів Інституту філології й одного професійного актора (один із записів представляв спонтанне мовлення). На основі вказаних звукозаписів було створено 42 файли анотацій у форматі ***.eaf**, що містять орфографічний запис аудіотекстів із пунктуаційними знаками та спрощену кириличну алофонемну транскрипцію для кожного аудіозапису. Створення дворівневих анотацій для кожного аудіофайлу проходило в чотири етапи в режимах розмітки, сегментації й транскрипції в програмі ELAN [8]. Також уже розроблено таблицю відповідності кириличних і латинських символів для наступного транскрибування фонозаписів за стандартами Міжнародного фонетичного алфавіту (IPA).

Варто зазначити, що різноманіття створених на сьогодні корпусів не обмежується вищевказаною інформацією, арсенал корпусних студій постійно доповнюється новими працями. Аналіз сучасного стану корпусних досліджень в Україні свідчить про те, що на сьогодні вже створено кілька корпусів писемного й усного українського мовлення, які становлять науковий інтерес і можуть слугувати матеріалом для лінгвістичних досліджень і наукових напрацювань. Водночас існує потреба у створенні більшої кількості анотованих баз даних літературного українського усного мовлення, що містять фонетичну розмітку.

**Перспективу нашого дослідження** вбачаємо в анотуванні аудіотекстів Корпусу усного транскрибованого українського мовлення за допомогою символів IPA, що дозволить, по-перше, узгодити Корпус із міжнародними фонетичними стандартами, а по-друге, використовувати цю базу даних із метою вивчення української мови як іноземної.

**Література:**

1. Бук С. Структурне анотування у корпусі текстів (на прикладі прози І. Франка) / С. Бук // Українська мова. – 2009. – № 3. – С. 59–71.

2. Васильєва Н. Б. Створення акустичного корпусу українського ефірного мовлення / Н. Б. Васильєва, В. В. Пилипенко, О. М. Радуцький та ін. // Оброблення сигналів і зображень та розпізнавання образів : праці 10-ї Всеукраїнської міжнародної конференції «УкрОбраз». – К., 2010. – С. 55–58.

3. Данилюк І. Корпус текстів для вивчення граматичної службовості / І. Данилюк // Лінгвістичні студії. – Донецьк : ДонНУ, 2013. – Вип. 26. – С. 224–229.

4. Демська О. Текстовий корпус: ідея іншої форми / О. Демська. – К. : ВПЦ НаУКМА, 2011. – 284 с.

5. Клименко Н. З комп'ютером – до глибин мови. З нагоди 80-річчя Інституту мовознавства ім. О. О. Потебні НАН України / Н. Клименко, Є. Карпіловська, Л. Кислюк // Світогляд. – 2010. – № 4. – С. 74–80.

6. Корпусні технології представлення мовного матеріалу [Електронний ресурс]. – Режим доступу : http://lexilab.oa.edu.ua/index.php/proekty/79-proekty/75-proekt-stvorennya-korpusiv-tekstiv.

7. Лінгвістичний портал mova.info [Електронний ресурс]. – Режим доступу : http://www.mova.info/.

8. Плахотнікова О. Використання програми ELAN в роботі зі звукозаписами корпусу українського усного мовлення / О. Плахотнікова // Українське мовознавство. – 2014. – Вип. 44. –Част. 1. – С. 238–243.

9. Робейко В. В. Моделирование особенностей спонтанной украинской речи в системах автоматического распознавания речевого сигнала / В. В. Робейко // Кибернетика и вычислительная техника. – 2012. – Вип. 170. – С. 76–85.

10. Сайт з розпізнавання та синтезу мовлення в Україні [Електронний ресурс]. – Режим доступу : speech.com.ua.

11. Український мовно-інформаційний фонд НАН України [Електронний ресурс]. – Режим доступу : http://lcorp.ulif.org.ua/LSlist/.

12. The Language Archive. ELAN. [Електронний ресурс]. – Режим доступу : https://tla.mpi.nl/tools/tla-tools/elan.

13. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. ELAN: a Professional Framework for Multimodality Research // Proceedings of the 5th International Conference on Language Resources and Evaluation. – LREC, 2006. – Pp. 1556–1559.