5. Хайчевська Т. М. Впливовий потенціал засобів організації тексту в сучасній французькій літературі / Т. М. Хайчевська // Мовні і концептуальні картини світу : зб. наук. пр. / КНУ ім. Т. Шевченка ; [відп. ред. О. І. Чередниченко]. – К. : ВПЦ «Київ. ун-т», 2010. – Вип. 29. – С. 331–335.
6. Frutiger A. L'homme et les signes. Signes, symboles, signaux / A. Frutiger. – P. : Atelier Perrousseaux, 2004. – 320 p.
7. Yveline B. Message media communication / B. Yveline. – Paris. : Magnard université, 1983. – 221 p.

Статтю подано до редколегії
29.03.2012 р.

УДК 81'373.72

І. М. Кульчицький – доцент, кандидат технічних наук кафедри прикладної лінгвістики Національного університету«Львівська політехніка»;
І. О. Ліхнякевич – старший викладач кафедри прикладної лінгвістики Національного університету «Львівська політехніка»;
Ю. О. Данчевська – викладач кафедри української та іноземних мов Львівського державного університету фізичної культури

Деякі аспекти використання лінгвістичних корпусів слов'янських мов

Роботу виконано на кафедрі прикладної лінгвістики НУ «Львівська політехніка» і на кафедрі української та іноземних мов ЛДУФК

У статті на основі аналізу друкованих та інтернет-джерел розглянуто деякі аспекти використання національних корпусів слов'янських мов.

Ключові слова: національний корпус, метарозмітка, морфологічна розмітка, синтаксична розмітка, доступ.

Кульчицкий И. М., Лихнякевич И. О., Данчевская Ю. О. Некоторые аспекты использования лингвистических корпусов славянских языков. На основе анализа печатных изданий и интернет-ресурсов в статье рассматриваются некоторые аспекты использования национальных корпусов славянских языков.

Ключевые слова: национальный корпус, метаразметка, морфологическая разметка, синтаксическая разметка, доступ.

Kulchytskyi I. M., Likhnyakevych I. O., Danchevska Yu. O. Some Aspects of Slavic Languages Linguistic Corpora Use. The article is dedicated to the review of some aspects of Slavic languages linguistic corpora use based on the analysis of print material and Internet-resources.

Key words: national corpus, metalinguistic tagging, morphological tagging, syntactic tagging, availability.

Постановка наукової проблеми та її значення. Під лінгвістичним корпусом, який є об'єктом корпусної лінгвістики, розумітимемо великий, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, поданий в електронному вигляді та призначений для різних мовознавчих студій [18, 3; 19, 10–13]. Зокрема такі корпуси широко використовує сучасна лексикографія [18, 11].

Свою історію лінгвістичні корпуси розпочали у 60-х роках XX ст. з появою Браунівського корпусу, розробленого в США В. Нельсоном Френсісом (W. Nelson Francis) та Г. Кучерою (H. Kucera), який став стандартом для створення наступних корпусів, серед яких: Ланкастерський корпус (Lancaster-Oslo-Bergen Corpus, LOB), Уппсальський корпус російської мови (The Uppsala Corpus of modern Russian), а також корпус усної англійської мови (London-Lund Corpus). Усі вони були невеликого обсягу (не більше мільйона слововживань), проте розкривали параметри відбору текстів для корпусу, а також розкривали коло потенційних питань, які можна вирішувати за їхньою

допомогою. Зокрема, на основі корпусів були створені корпуси для лексикографічних праць (American Heritage Intermediate), вивчення розмовної англійської (Lancaster / IBM Spoken English Corpus, Corpus of Spoken American English, тощо). Із розвитком комп'ютерних технологій у 80–90-х почався і новий етап розробок корпусів обсягом понад 100 мільйонів слововживань (т. з. друге покоління корпусів) – з'являється Британський Національний Корпус (British National Corpus), Американський Національний Корпус (American National Corpus), Корпус Сучасної американської англійської (COCA), започатковано так званий проект моніторингового корпусу – The Cobuild Project / The Bank of English (під керівництвом Дж. Сінклера, Університет Бірмінгема), який щорічно поповнюється новими слововживаннями і є основою для створення словників видавництва Harper Collins, комерційні проекти корпусів – спільний проект Ланкастерського Університету та видавництва Pearson Longman (The Longman Corpus Network), міжнародні проекти – The International English Corpora (20 паралельних підкорпусів варіантів англійської мови тих країн, де вона є офіційною або другою офіційною мовою), Gigaword Corpora (публіцистика та новини англійською, арабською, китайською та іншими мовами) [2: 3; 4]. Поява паралельних корпусів сприяє розвиткові перекладознавства, удосконалення машинного перекладу, вивченню іноземних мов тощо [21]. Загалом розвиток корпусної лінгвістики сприяє і розвиткові сфер застосування її здобутків – від отримання різноманітних статистичних даних (конкордансів, частотності слів тощо), розробки лексикографічних праць, удосконалення лінгводидактчних матеріалів (граматик та підручників), вивчення текстів та усного мовлення для визначення функціональних стилів, соціолінгвістики, до перевірки автентичності текстів чи навіть для використання у розслідуванні злочинів (forensic linguistic analysis) [23, 578–590; 24].

Щоб використовувати корпус для такого широкого спектру завдань (які не завжди є лінгвістичними), важливим є не стільки велика кількість текстових даних у корпусі, а їх репрезентативність, тобто достатнє та пропорційне представлення в корпусі текстів різних періодів, жанрів, стилів, авторів тощо. Найважливішим для лінгвістичних досліджень в корпусі є наявність морфологічної, синтаксичної, семантичної розміток. Проте далеко не всі з наявних корпусів задовольняють ці вимоги. Більшість із них можуть представити лише метарозмітку (екстралінгвістичну інформацію про тексти корпусу), морфологічну, рідше – синтаксичну та семантичну чи інші типи розміток (анафоричну чи просодичну) [18, 10–11]. Все це відбувається внаслідок відсутності або недоліків розроблених програм для автоматичного анотування текстів корпусів, внаслідок чого їх потрібно повторно вручну чи напівавтоматично виверяти та виправляти, що не є швидким процесом, враховуючи сучасні обсяги корпусів. **Актуальність** дослідження полягає в тому, що корпусну лінгвістику вважають порівняно молодою дисципліною, розвиток якої відбувається нерівномірно в різних країнах. **Метою** цієї статті є аналіз літератури та Інтернет-джерел для огляду сучасного стану національних корпусів слов'янських мов. Загалом питанням стану національних корпусів слов'янських мов цікавились та досліджували такі вчені, як Т. І. Рєзникова та В. П. Захаров [18; 25–32; 20].

Виклад основного матеріалу та обґрунтування отриманих результатів дослідження. Майже всі слов'янські мови мають свої національні корпуси або розпочали роботу над ними. У групі західнослов'янських мов варто виокремити корпус чеської мови (ЧНК), розробка якого розпочалася в 90-і роки і який створювався на основі Браунівського корпусу. Саме для чеського корпусу, уперше для слов'янських мов, був розроблений корпус із синтаксичним анотуванням – *Prague Dependency Treebank.* До складу корпусу входить діахронічна та синхронічна частини, остання містить підкорпуси письмової мови, усної мови та публіцистики; підбір текстових даних збалансований, оскільки розробники проводили соціолінгвістичні дослідження, згідно з якими включали у корпус ті чи інші тексти. Щодо типів розмітки, наявних у чеському корпусі, то тут представлена метарозмітка, морфологічна розмітка, виконана автоматично з використанням методів зняття граматичної омонімії. Проте морфологічна розмітка відсутня у підкорпусах усного мовлення. Синтаксична, семантична чи інші типи розміток відсутні взагалі. Пошук у корпусі здійснюється з допомогою спеціально розробленої програми (т. з. корпус-менеджера) Bonito, яка знаходить слова за словоформою, лексемою (у корпусах з лематизацією), послідовністю словоформ і граматичними ознаками (у корпусах із морфологічною розміткою). Доступ до усіх функціональних можливостей корпусу можна отримати, зареєструвавшись (безкоштовно доступ буде надано, якщо корпус використовуватиметься для дослідницьких цілей). Доступ без реєстрації обмежений отриманням 50 контекстів із зазначенням кількості всіх знайдених у корпусі прикладів слововживань [5].

Словацький національний корпус почали створювати 2002 р., і на сьогодні він представляє синхронний корпус текстів із 1955 р. корпус усного мовлення відсутній. У корпусі здійснена метарозмітка (з ширшим набором параметрів, ніж у ЧНК), морфологічна розмітка (автоматична зі знятою омонімією, половина корпусу розмічена вручну), інші типи розміток відсутні взагалі. Пошук у корпусі, як і у ЧНК, здійснюється за допомогою програми Bonito за такими ж параметрами, як і у ЧНК. Доступ до всього корпусу можливий після реєстрації, безкоштовно ж – у разі використання корпусу для дослідницьких цілей. Доступ без реєстрації обмежений, як і у ЧНК [6].

Національний корпус польської мови складається з трьох окремих корпусів – IPI PAN, PELCRA та PWN – і перебуває у вільному доступі в мережі Інтернет [7]. Перший з них – IPI PAN – характеризується лематизацією та автоматичною морфологічною розміткою, що найбільш підходить під визначення сучасних типів корпусів. Водночас підкреслено не репрезентативність зібраних у корпусі текстових даних. Для пошуку в корпусі розроблено спеціальний корпус-менеджер Poliqarp, який здійснює пошук за тими ж параметрами, як і програма для ЧНК та СНК [22, 60–81]. Корпус PELCRA створений за моделлю Британського національного корпусу з 1996 року як спільний проект університетів у Лодзі та Ланкастері. Тут одночасно із створенням одномовного корпусу проведено роботу над створенням паралельного польсько-англійського корпусу та польського навчального корпусу англійської мови. До складу корпусу входить 90 % письмових та 10 % усних текстів. Корпус має метарозмітку, проте з обмеженим набором параметрів, морфологічна та інші типи розмітки відсутні взагалі. Так, пошук по корпусу може здійснюватися лише для отримання статистичних даних.

Корпус PWN розроблений Польським науковим видавництвом і використовується ним для створення словників. У корпус увійшли тексти різних типів, а також публіцистика. На відміну від стандартного для корпусів методу включення повних текстів, PWN містить лише їх фрагменти. Метарозмітка включає стандартні параметри, крім того наявне тегування в текстах слів іншомовного походження, діалектизмів, неправильних форм слововживань (з подачею правильної форми). Здійснена лематизація, без зняття омонімії, проте морфологічна розмітка відсутня. Безкоштовно доступна лише частина корпусу з обмеженими можливостями пошуку [8].

Серед південнослов'янських мов найрозвиненішим вважають корпус словенської мови. Корпус FIDA та його розширена версія Fida PLUS відрізняють кількісно, проте вважаються збалансованими і мають мета- та лінгвістичну розмітки. Граматична омонімія знята частково. Пошук у корпусі можна здійснювати за тими ж параметрами, що і у ЧНК чи ХНК. Доступним є корпус Fida PLUS, проте для безпосереднього доступу рекомендовано реєструватись. Із 2003 року ведуть роботу над створенням синтаксично розміченого корпусу Slovenian Dependency Treebank. Для доступу до цього типу корпусу потрібно звернутися з листом до розробників [9].

Хорватський національний корпус характеризується незбалансованістю зібраних текстів (97 млн слововживань із газет та журналів та більше 4 млн слововживань – із художньої літератури), наявністю метарозмітки і часткової лематизації та морфологічної розмітки (на незначному обсязі газетних текстів). Доступ до корпусу абсолютно відкритий і вільний, проте його можливості обмежені [10].

Корпус сербської мови складений на основі вибірки з 11 млн слів і охоплює період з XII-го століття до наших днів. Кожне слово вручну морфологічно анотоване, з кількістю графем, складів та фонологічною структурою включно. У тексті зроблено структурну розмітку – початок та кінець речення, абзаци. В Інтернеті корпус недоступний (лише зразки розмітки обсягом по 500 словоформ кожного з п'яти підкорпусів) [11].

Болгарська мова не має власного корпусу, проте з 2009 року ведуться роботи з його створення. Болгарський національний корпус створюється в Інституті болгарської мови на базі двох його департаментів – комп'ютерної лінгвістики і болгарської лексикології та лексикографії. Він складається з кількох окремих електронних корпусів, розроблених протягом останніх десяти років. Сьогодні болгарський національний корпус складається з 400 млн слів і містить понад 11 000 текстів. Матеріали, включені в корпус, відображають стан болгарської мови (в основному в письмовій формі) з 1945 р. і до сьогодні. Окремі частини містять детальний уніфікований опис, що полегшує їх подальшу обробку та групування. Тексти корпусу класифікують залежно від їх належності до певної тематичної групи (наприклад, художня література або публіцистика). Болгарський національний корпус склада-

ється з трьох великих корпусів і 13 менших, створених для конкретних цілей. Розробка здійснювалась на основі Браунівського корпусу, проте з деякими відмінностями. Здійснено автоматичну морфологічну розмітку. Доступ до корпусу вільний, обмежений для незареєстрованих користувачів [12]. Проводять роботи зі створення синтаксично анотованого корпусу – BulTreeBank – в Лабораторії лінгвістичного моделювання при Болгарській академії наук. Він представлений у вільному доступі, проте у форматі XML файлу.

Стосовно македонської мови, сьогодні є лише доступним т. з. «Архів македонської мови», який має стати основою майбутнього корпусу. Крім того, є Македонський корпус, розроблений Лабораторією «Текст» при Університеті Осло, доступ до якого можливий за умов реєстрації [13].

Якщо говорити про східнослов'янські мови, то білоруська представлена доступним в Інтернеті корпусом наукових текстів, розробленим при Білоруському національному технічному університеті, проте жодної інформації про склад та історію створення не подано. Увесь інтерфейс представлений білоруською мовою [14].

Найрозвиненішим є корпус російської мови, робота над яким розпочалася у 2001 р. (правда, до цього вже існували розробки корпусів російської мови, проте створені за кордоном – Уппсальський корпус сучасної російської мови, Тюбінгенський корпус, ХАНКО) [20]. На сьогодні національний корпус російської мови (НКРЯ) налічує більше 300 млн слововживань, містить десять підкорпусів (окрім основного, є також мультимедійний, усного мовлення, акцентологічний (з просодичною розміткою), поетичний, діалектний, газетних текстів, паралельний та синтаксично анотований). Здійснені всі типи розмітки, не на всьому об'ємі корпусу, проте проект і далі розвивається. Відповідно пошук у корпусі може здійснюватися майже за всіма параметрами: словоформа, лексема, послідовність словоформ, граматичні ознаки, синтаксичні структури та семантичні ознаки. Доступ до корпусу відкритий і вільний [15].

Тестова версія корпусу текстів української мови, розроблена співробітниками лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка під керівництвом Н. П. Дарчук, доступна в мережі Інтернет. За даними, доступними на сайті проекту, дослідницький корпус сучасної української мови обсягом у 3 млн словоформ побудований як інформаційно-довідкова система. Це тексти в електронній формі, призначені не для читання, а для з'ясування різних питань, пов'язаних з українською мовою. Пошук можна здійснювати за типом підкорпусу (художня проза, нехудожні твори, поетичні або фольклорні тексти), словом та/або його морфологічними характеристиками, а також додатковим параметром є стать автора. Результат пошуку подано у вигляді таблиці, де зліва – шукане слово в контексті, виділене жирним шрифтом, а справа – посилання на джерело із екстралінгвістичною інформацією [16; 1, 46–49].

Крім того, існує корпус, який розробляє мовно-інформаційний фонд НАН України під керівництвом В. А. Широкова, його використову фонд для створення різноманітних електронних словників [17]. Проте для широкого загалу цей корпус недоступний.

Висновки. Отже, у підсумку можна зазначити, що, незважаючи на існування корпусів, вони не завжди відповідають усім стандартам, притаманним справжнім національним корпусам, приміром у плані репрезентативності, наявності різних типів розмітки. Отже, варто зазначити потребу подальшого розвитку та вдосконалення наявних корпусів, а також появи нових, які перебувають на стадії розробки. Крім того, іншою проблемою можна вважати стандартизацію методів розмітки, яка відрізняється в різних корпусах, а також відмінності в розробках інтерфейсу корпус-менеджерів для користувачів – одні з них розроблені для недосвідчених користувачів (наприклад, НКРЯ), інші, навпаки, для користувачів, обізнаних з правилами задавання запитів, атрибутами, які використовувались при розмітці текстів корпусу тощо (наприклад, ЧНК, ХНК, СНК). Отже огляд сучасного стану корпусів слов'янських мов дозволяє охарактеризувати основні методики їх розробки, а також окреслити перспективи розвитку корпусів загалом.

Список використаної літератури

1. Дарчук Н. Дослідницький корпус української мови: основні засади і перспективи / Н. Дарчук // Вісн. Київського нац. ун-ту ім. Тараса Шевченка. Серія : «Літературознавство. Мовознавство. Фольклористика». – 2010. – № 21. – С. 45–49.
2. [Електронний ресурс]. – Режим доступу : http://www.pearsonlongman.com/dictionaries/corpus/index.html.

3. [Електронний ресурс]. – Режим доступу: http://ice-corpora.net/ice/index.htm.
4. [Електронний ресурс]. – Режим доступу: http://projects.ldc.upenn.edu/.
5. [Електронний ресурс]. – Режим доступу : http ://ucnk.ff.cuni.cz/english/index.php.
6. [Електронний ресурс]. – Режим доступу : http ://korpus.juls.savba.sk/index_en.html.
7. [Електронний ресурс]. – Режим доступу : http ://nkjp.pl/index.php?page=6&lang=1.
8. [Електронний ресурс]. – Режим доступу: http ://korpus.pwn.pl/index_en.php.
9. [Електронний ресурс]. – Режим доступу : http ://www.fidaplus.net/Info/Info_index_eng.html.
10. [Електронний ресурс]. – Режим доступу:http ://www.hnk.ffzg.hr/cnc.htm.
11. [Електронний ресурс]. – Режим доступу : http ://www.serbian-corpus.edu.rs/ns/preview/preview.htm.
12. [Електронний ресурс]. – Режим доступу : http ://www.sciencenewsline.com/technology/20100527000070 06.html.
13. [Електронний ресурс]. – Режим доступу : http ://omilia.uio.no/swamp/index.php.
14. [Електронний ресурс]. – Режим доступу : http ://grid.bntu.by/corpus/index.php?word1=&numw=1&scope= sent&wdist=0.
15. [Електронний ресурс]. – Режим доступу : http ://ruscorpora.ru/index.html.
16. [Електронний ресурс]. – Режим доступу: http ://www.mova.info/corpus.aspx?l1=209.
17. [Електронний ресурс]. – Режим доступу: http ://lcorp.ulif.org.ua/dictua/.
18. Захаров В. Корпусная лингвистика : учебно-метод. Пособие / В. Захаров. – СПб. : [б. и.], 2005. – 48 с.
19. Корпусна лінгвістика : монографія / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна, О. М. Костишин, М. Ю. Кригін ; НАН України, Укр. мов.-інформ. фонд. – К. : Довіра, 2005. – 472 с.
20. Резникова Т. Славянская корпусная лингвистика: современное состояние ресурсов / Т. Резникова // Нац. корпус русского яз. : 2006–2008. Новые результаты и перспективы. – СПб. : Нестор-История, 2009. – С. 402–461.
21. Соснина Е. П. Корпусная лингвистика и корпусный подход в обучении иностранному языку / Е. П. Соснина // Corpus Linguistics and Corpus-Based Approach in Foreign Language Teaching [Електронный ресурс]. – Режим доступа : http://ling.ulstu.ru/linguistics/resourses/literature/articles/corpus_education_translation/.
22. Adam Przepiórkowski The IPI PAN Corpus preliminary version. [Electronic resource]. – Access mode : http ://nlp.ipipan.waw.pl/~adamp/Papers/2004-corpus/book_en.pdf.
23. [O'Keeffe A. The Routledge Handbook of Corpus Linguistics // O'Keeffe A., McCarthy M. – New York, The Routeledge, 2010. – 711 p.
24. Ukrainian National Linguistic Corpus and its application [Electronic resource] / Shyrokov V., Bugakov O., Krygin M., Sydorchuk N. // SlaviCorp 2011 Conference Presentations. – Access mode : http ://www. slavicorp.polon.uw.edu.pl/streszczenia.html.

Статтю подано до редколегії
14.03.2012 р.

УДК 811.161.2'373

Лю Фейхун – аспирант Днепропетровского национального университета имени Олеся Гончара

Эмоциональное поле междометий как проблема лексикографического представления

Работа выполнена на кафедре перевода и лингвистической подготовки иностранцев ДНУ им. Олеся Гончара

У статті розглянуто види вигуків згідно з їх значенням та засоби їх тлумачення у словнику.
Ключові слова: вигук, словник, тлумачення, емоція.

Фейхун Лю. Эмоциональное поле междометий как проблема лексикографического представления.
В статье рассматриваются виды междометий в соответствии с их значением и способы их толкования в словаре.
Ключевые слова: междометие, словарь, толкование, эмоция.