ХРОНІКА

1–3 жовтня 2019 р. у місті Синтра (Португалія) відбулася VI Міжнародна конференція з електронної лексикографії «eLex 2019: Electronic lexicography in the 21st century (Smart Lexicography)», організатором якої був Університет Коїмбра. У ній узяли участь науковці з тридцяти країн світу: Австрії, Бельгії, Болгарії, Бразилії, Великобританії, Данії, Естонії, Ізраїлю, Ірландії, Іспанії, Італії, Латвії, Мексики, Нідерландів, Німеччини, Польщі, Португалії, Росії, Сербії, Словаччини, Словенії, США, України, Фінляндії, Франції, Хорватії, Чехії, Чилі, Швеції та ін. Перша конференція відбулася 2009 р. в Бельгії, а згодом стала проводитися через кожні два роки. Її засновники — Європейська асоціація лексикографів EURALEX та спеціальна дослідницька група SIGLEX — ставили за мету обговорення останніх напрацювань і досягнень в електронному словникарстві, обмін думками з питань, що становлять інтерес для людей, які працюють у лексикографії та суміжних сферах.

Робоча програма конференції містила 5 ключових доповідей, 18 секцій, 49 усних доповідей, 24 стенди, 18 демо-презентацій, одну спонсорну сесію та дві церемонії вручення премій (перша — премії імені Адама Кілгаріффа, друга — премії від фонду Hornby Educational Trust).

Словосполучення *smart lexicography* у назві конференції свідчить про зосередження науковців на технологічних аспектах електронного словникарства. Один з них стосується адаптації різних лексикографічних праць до цифрових форматів. Так, в останнє десятиріччя використання мобільних пристроїв (смартфонів) значно зросло Проте більшість словників досі не пристосована до таких цифрових форматів, і це потребує вирішення нових завдань у цьому напрямі, адже більшість користувачів використовує мобільні пристрої щодня. Іншим аспектом є технології багаторазового використання змісту словника. Саме словники часто залишаються ізольованими об'єктами, тоді як потреби та звички користувача вказують на те, що було б набагато корисніше пов'язати їх з іншими словниками і мовними ресурсами або навіть інтегрувати в різні інструменти. Усе це створює, зввичайно, численні проблеми, особливо якщо ці електронні ресурси належать різним видавництвам з різних країн. Такі самі проблеми спостерігаються у випадку повторного використання змісту виданих словників під час створення нових ресурсів. З огляду на це важливим постає забезпечення сумісності форматів репрезентації словників, а також розроблення API-сервісів та інструментів конвертації, що сприяють цій сумісності, а також наближають лексикографію до семантичної мережі.

Напередодні конференції 30 вересня пройшов семінар «Collocations in Lexicography: existing solutions and future challenges» (Колокації в лексикографії: наявні рішення та майбутні завдання) під керівництвом Ізтока Косема, провідного співробітника Інституту прикладних словенських досліджень (м. Любляна, Словенія). У роботі семінару взяли участь провідні фахівці з комп'ютерної лінгвістики: Володимир Бенко (Словенія), Мілош Якубічек (Чехія), Ракел Амаро (Португалія), Жиль Моріс (Бельгія) та ін. Мета семінару — розглянути й обговорити теоретичні та практичні питання,

пов'язані з лексикографічним упорядкуванням словосполучень. Коло питань, висвітлених у доповідях, охоплювали:

1) створення та використання корпусів як джерел колокацій, маркування й екстракція колокацій з корпусів; статистичні дослідження для визначення найчастотніших сполучень слів;

2) синтаксис і семантику колокацій: семантичні зв'язки між колокаціями та їх подання в електронних словниках; граматичний формалізм для кодування колокацій у лексикографічній базі даних та для автоматичної екстракції з корпусів;

3) проблему відокремлення колокацій від інших багатослівних виразів у завданнях автоматичного опрацювання текстів;

4) упорядкування колокацій в електронних словниках відповідно до потреб користувача та побудову дефініцій до колокацій. Оскільки словники виконують не лише довідкову, а й навчальну функцію, робота з колокаціями дає змогу користувачеві відпрацьовувати навички самостійної роботи з опанування мови і розвивати комунікативну компетенцію. Дослідження колокацій проводилися на матеріалі не лише європейських мов (словенська, чеська, російська, португальська), а й тих, що на сьогодні залишаються мало дослідженими, зокрема мова народності сога, одна з мов банту, розповсюджена в Уганді.

Урочисте відкриття конференції (1 жовтня 2019 р.) почалося з привітання голови організаційного комітету Танари Кун, доктора філософії з прикладної лінгвістики, наукового співробітника центру загального та прикладного мовознавства університету Коїмбра. У своїй промові вона виокремила актуальні напрями сучасної електронної лексикографії та ознайомила учасників з роботою центру, а також розповіла про його тісну співпрацю з Міжнародним інститутом португальської мови у виробленні норм та принципів укладання словників португальської мови. Серед головних досягнень центру вона відзначила створення великого інтернет-порталу португальської мови (http://www.portaldalinguaportuguesa.org/) — платформи, що містить серію лінгвістичних словників та довідкових матеріалів з правопису. Розроблений ресурс розрахований на широке коло користувачів, у тому числі для науковців, які досліджують португальську мову в різних її аспектах. Контент, розміщений на порталі, є у вільному доступі і постійно переглядається та оновлюється.

Як уже зазначалося, під час церемонії відкриття конференції Ізток Косем вручив премії від фонду Hornby Educational Trust. Цей фонд 1961 р. створив Альберт Синді Хорнбі, англійський лексикограф, відомий як укладач Оксфордського словника. Основною діяльністю фонду є видача грантів закордонним викладачам англійської мови для підвищення ними своєї професійної кваліфікації у Великобританії. Докладну інформацію про фонд та умови отримання грантів можна знайти на сайті фонду https://www.hornby-trust.org.uk/.

Маргарита Коррейа, доктор філософії Лісабонського університету в галузі португальської лінгвістики, у своїй доповіді запропонувала обговорити тему створення електронного орфографічного словника португальської мови (VOC). Як відомо, португальська — це плюрицентрична мова, оскільки нею розмовляють у восьми країнах на чотирьох континентах. Із 1911 року в ній діяли дві різні орфографічні норми: бразильська і португальська. Протягом XX ст. уряди Португалії

та Бразилії прагнули отримати набір правил правопису, що охоплювали б обидва національних варіанти з використанням системи правопису, яка є в основному фонематичною. Починаючи з 1990 р., всі португаломовні країни пов'язані між собою Орфографічною угодою з португальської мови (AOLP 90). Однак протягом понад двох десятиліть цей набір правил не вдалося сформулювати у вигляді загального словника правопису для офіційного використання у всіх цих країнах. У зв'язку з цим постала необхідність створення VOC — Vocabulário Ortográfico Comum da Língua Portuguesa — Загального орфографічного словника португальської мови. Його цифрова платформа інтегрує словники правопису кожної португаломовної країни, укладені відповідно до загальних орфографічних норм і принципів лексикографічної репрезентації. У доповіді виокремлено головні питання створення такого словника, а також підкреслено його політичну роль і значення (режим доступу до словника: http://iilp.cplp.org/voc/).

Робота конференції розпочалася у трьох секціях: «Програми-коректори, користувачі та проектування словників», «Історичні словники, етимологія та інтеграція», «Малоресурсні мови та термінологія». У першій секції доповіді були присвячені проблемам використання штучного інтелекту у програмах-коректорах. У другій секції предметом обговорення стали технології інтеграції історичного словника польської мови з корпусом текстів XVII–XVIII ст., створення генеративного етимологічного словника індоєвропейських мов, а також використання семантичних мереж у дослідженні еволюції мови. Третя секція була присвячена проблемам створення електронних лексикографічних ресурсів для малоресурсних мов (наприклад, тібетської, санскриту), тобто мов, для яких ще не розроблено великих одномовних та паралельних корпусів текстів.

Зростання інтересу до автоматизованої підтримки процесу написання тексту, а також останні досягнення у сфері інформаційних технологій сприяли розробленню нового покоління програм-коректорів. Більшість лексикографів (А. Франкенберг-Гарсія, С. Грангер, С. Тарп, Л. Ваннер та ін.) розглядають програми-коректори як засіб для навчання мови. Такий інструмент призначено користувачам, що не мають навичок або мають слабкі навички роботи зі словниками. Вважається, що програми-коректори прискорюють процес засвоєння мови і можуть бути альтернативою словникам. На особливу увагу в цьому контексті заслуговує доповідь «ColloCaid: Assisting Writers with Academic English Collocations» (ColloCaid: помічник для письменників у творенні англійських літературномовних колокацій), автори А. Франкенберг-Гарсія, Р. Лью, Ж. Пол Різ та ін., у якій ідеться про проєкт ColloCaid (режим доступу: https://collocaid.uk/prototype/editor/public/). Проєкт має на меті створення інструменту для редагування текстів, який би допомагав вибирати слова, що добре сполучаються одне з одним. Відмінною рисою ColloCaid є те, що він не обмежується наданням довідки з колокацій. Основна його мета — підвищувати в реальному часі обізнаність про колокації, яких автори можуть не пам'ятати або не знати, не перериваючи при цьому процесу написання текстів.

Словникові мережі є цікавими мовними інструментами й хорошими довідниками для тих, хто вивчає тематичні слова та їх еволюцію в тій чи іншій мові. Використовуючи технології семантичних мереж, тобто математичних структур, що репрезентують зв'язки між словами, можна

відслідкувати еволюцію розвитку мови. Іспанські дослідники Каміло Гаррідо, Клаудіо Гутьєрес та Гільєрмо Сото у своїй доповіді «The Semantic Network of the Spanish Dictionary during the last century: structural stability and resilience» (Семантична мережа іспанського словника протягом минулого століття: структурна стабільність та стійкість) за об'єкт дослідження взяли семантичну мережу іспанського словника, яка охоплювала період з 1925 по 2014 рр. Дослідники виявили, що глобальні структурні властивості семантичної мережі іспанського словника надзвичайно стабільні. З іншого боку, з роками локальні властивості змінюються, пропонуючи уявлення про еволюцію лексикону.

Як показує досвід польських лексикографів (Олександра Вечорек та ін.), висвітлений у доповіді «Integration of the Electronic Dictionary of the 17th−18th c. Polish and the Electronic Corpus of the 17th and 18th c. Polish Texts» (Інтеграція електронного словника польської мови XVII–XVIII ст. із корпусом польських текстів XVII–XVIII ст.), інтеграція історичного словника з історичним корпусом у поєднанні електронного словника польської мови з корпусом текстів XVII–XVIII ст. («Корпусом бароко») надає багато можливостей як для лексикографів, так і для користувачів. По-перше, інтеграція дає змогу ідентифікувати відсутні словникові записи (тобто ті, що існують у корпусі, але відсутні у словнику). По-друге, пошук прикладів використання для існуючих запитів спрощено. По-третє, стала можливою перехресна перевірка знаходження повного набору словоформ. І, нарешті, спрощено пошук фразових дієслів, прислів'їв тощо (усіх наявних у словникових записах).

Будь-яке порівняльно-історичне дослідження потребує відповідних програмних засобів. До таких належить протоіндоєвропейський лексикон (скорочено PIELex). Це генеративний етимологічний словник індоєвропейських мов (режим доступу: http://pielexicon.hum.helsinki.fi) — перший у світі словник, здатний автоматично створювати лексичні корені для понад 120 архаїчних мов індоєвропейської сім'ї. Крім цього, для здійснення зворотного процесу розпочато роботу над програмами механічної генерації протоіндоєвропейських словоформ з використанням даних з індоєвропейських мов. Функціонування словника PIELex забезпечують алгоритми опрацювання даних (починаючи із семантики й закінчуючи морфологією) та вихідна структура самої протоіндоєвропейської мови.

Другий день конференції розпочався виступом Олександра Гейкена (Берлін, ФРН) «Центр цифрової лексикографії німецької мови: нові перспективи інтелектуальної лексикографії». Головною діяльністю центру є забезпечення комплексного та емпірично достовірного опису німецької мови від її витоків до сьогодення. Для цього чотири німецькі академії в Берліні (координатор), Геттінгені, Лейпцигу і Майнці об'єднали свої зусилля. Доповідач зазначив, що в академіях накопичено багатий досвід створення різних словникових проєктів, що охоплюють як історичні, так і сучасні словники, у тому числі Grimmsches Wörterbuch, словники давньонімецької, середньонімецької, ранньонімецької і цифровий словник (DWDS) сучасної німецької мови. Крім цього, Центр співпрацює з Інститутом німецької мови ім. Лейбніца (IDS) в галузі неологізмів і сучасних текстових корпорацій. Щоб забезпечити повсюдний пошуковий інтерфейс для цих різноманітних джерел словників, у найближчі роки буде потрібен значний обсяг роботи з інтеграції, включаючи роботу над спільними форматами, списками лем, а також перехресні посилання зі словників на корпуси.

Далі свою роботу розпочали стендова та демонстративна секції. На стендовій секції представлено 24 доповіді різної тематичної спрямованості, починаючи з корпусних технологій і закінчуючи окремими проблемами укладання електронних словників загальномовної та термінологічної лексики, зокрема:

1) добір речень для створення навчальних конкордансів, що показують особливості вживання того чи іншого слова або фрази в реальному мовленні автентичними носіями мови;

2) екстракція термінів та дефініцій із корпусів текстів; використання семантично анотованих корпусів для добору ілюстрацій до тлумачного словника;

3) автоматизована екстракція колокацій, проблеми лексикографічного опису колокацій та опис сполучуваності слів у словнику;

4) автоматична побудова багатомовного словника колокацій з використанням дистрибутивної семантики;

5) створення інструментів для автоматичної побудови дефініцій;

6) екстракція термінів з використанням формул дефініцій та прив'язування кожної формули до конкретного семантичного відношення;

7) автоматичне створення гіперпосилань, що зв'язують слово в дефініції з іншими реєстровими словами;

8) побудова графічного інтерфейсу для електронних словників загальномовної та спеціальної лексики.

На демонстраційній секції всім учасникам конференції надали можливість ознайомитися з різними програмними засобами: електронними словниками та системами підтримки лексикографічного процесу (докладну інформацію про ці лексикографічні інструменти та ресурси можна отримати тут: https://indico2.conference4me.psnc.pl/event/12/sessions/207/#20191002). У роботі секції взяли активну участь представники Sketch Engine. Ця корпорація спеціалізується на створенні інструменту (SkELL), що надає студентам і викладачам інформацію щодо уживання конкретної фрази або слова в усному мовленні носіїв тієї чи іншої мови. Такий інструмент працює з англійською, німецькою, італійською, чеською, російською та естонською мовами. SkELL доступний на сайті корпорації: https://www.sketchengine.eu/skell/. Інструмент відображає: 1) конкорданс, що подає приклади вживання шуканого слова в мовленні; 2) колокації з конкретним словом та 3) синоніми. Усі приклади, словосполучення та синоніми взято з багатомільйонних текстових вибірок автоматично, без використання ручної праці.

Другий день конференції завершився роботою секцій «Автоматична / автоматизована лексикографія», «Розумні словники, інструменти та нові підходи» і «Онтології та зв'язані лексичні дані». На першій обговорювали питання створення інструменту для автоматизованого вилучення лексем певного типу (зокрема, неологізмів), автоматизацію створення багатомовного словника на основі великого корпусу та побудову онтології на основі моделі OntoLex-lemon. На другій секції були репрезентовані нові підходи до використання можливостей інтернету у розширенні контенту словників, методи і прийоми автоматичного івідстежування життєвого циклу лексем з використанням діахронічного корпусу та інтеграція етимологічного словника з мережею лінгвістичних зв'язаних даних (Linguistic Linked Open Data network). Третя секція була присвячена створенню проєктів динамічних баз знань певної галузі (EcoLexicon: https://ecolexicon.ugr.es/visual/index_

en.html) та систем лексичних зв'язаних даних різних ресурсів і країн для дослідження мов певного регіону (VerbaAlplina: https://www.verba-alpina.gwi.uni-muenchen.de/).

У конференції вперше взяли участь представники з України — директор Українського мовно-інформаційного фонду НАН України, академік В. А. Широков, старший науковий співробітник Українського мовно-інформаційного фонду НАН України І. В. Остапова та доцент кафедри інтелектуальних комп'ютерних систем Національного технічного університету «Харківський політехнічний інститут» Є. В. Купріянов. На стендовій секції вони представили доповідь «Лексикографічна система тлумачного словника у цифровому середовищі», де наочно продемонстрували досвід Фонду в створенні цифрових словників, розробленні технологій їхньої інтеграції, а також побудові інструментів для роботи з ними. Зокрема, було репрезентовано тлумачний, етимологічний і граматичний словники української мови. Як останню розробку було показано програмний комплекс для аналізу тексту цифрового словника іспанської мови. Основними перевагами цифрових словників, створених у фонді, є: практично необмежений потенціал інтеграції різних лінгвістичних фактів в одному об'єкті, здатність відбивати мовну динаміку, ефективність навігації по структурних елементах, можливість проведення обчислювальних експериментів. Чимале значення має можливість багаторазового використання одного разу сформованих у цифровому середовищі лексикографічних структур і масивів багатьма професіоналами — лінгвістами, лінгвотехнологами і видавцями.

Останній день конференції був присвячений обговоренню окремих питань створення електронних словників, онтологій та тезаурусів. Було констатовано, що конференція пройшла на високому рівні, організатори дуже якісно спланували програму заходу та надали чудову можливість усім учасникам поділитися власними досягненнями та досвідом у галузі електронної лексикографії. Усі повні публікації та тези доповідей розміщено на сайті конференції за адресою: https://elex.link/elex2019/proceedings-download/.

Головною темою наступної конференції eLex 2021, яка відбудеться 2021 р. у місті Брно (Чехія), буде «Лексикографія постредагування». Цифрова лексикографія за останні 30 років досягла настільки великого прогресу, що багато елементів словника (наприклад, реєстр слів, словозміна, переклади або ілюстративні приклади) можна автоматично створювати на базі досить великого й анотованого текстового корпусу. Цей технічний прогрес веде до появи нових методологічних підходів, за яких велика частина редакційної роботи складається з постредагування автоматично створеного контенту — аналогічно до постредагування машинно перекладених текстів. Ці зміни сприятимуть прискоренню редакційної роботи і дозволять редакторам зосередитися на найскладніших питаннях, але наразі головною проблемою залишається впровадження та реалізація цих змін у парадигмі сформованих лексикографічних традицій. Докладну інформацію про eLex 2021 можна прочитати на сайті: https://elex.link/elex2021/.

Є. КУПРІЯНОВ
(Харків)

І. ОСТАПОВА
(Київ)