УДК 81'33

КОМП'ЮТЕРНА ЛІНГВІСТИКА І WEB 2.0

Кузьменко Дмитро Федорович
Київський національний університет імені Тараса Шевченка

Теоретичні дослідження з комп'ютерної лінгвістики та їхнє прикладне втілення є обов'язковим елементом більшості сучасних Інтернет-проектів. У статті йдеться про роль прикладної лінгвістичної складової у принципово нових мережевих концептуальних та технологічних явищах, що умовно називаються мережею другого покоління або Web 2.0.

Ключові слова: *комп'ютерна лінгвістика, прикладна лінгвістика, Web 2.0, Інтернет, інформаційні технології, обробка природної мови.*

Комп'ютерна лінгвістика відіграє суттєву роль у всіх сферах розвитку та функціонування сучасних інформаційних технологій, зокрема в мережі Інтернет. Комунікативний аспект Інтернету є визначальним, а основним засобом комунікації була і залишається природна мова, тому автоматичне й автоматизоване комп'ютерне опрацювання природної мови є складовою більшості сучасних мережевих інформаційних технологій. Для визначення найбільш актуальних напрямків досліджень з прикладної лінгвістики у цій сфері на даний момент корисними є загальні огляди найновіших мережевих тенденцій та роль у них прикладних лінгвістичних компонентів. Подібних досліджень в українському прикладному мовознавстві ще не було, а Інтернет розвивається надзвичайно динамічно і швидко, тому є велика потреба у всебічному розгляді даного питання. Ми не претендуємо на вичерпність висвітлення теми, а маємо за мету звернути на неї увагу й охарактеризувати її у загальних рисах.

На початку нового тисячоліття у мережевих концепціях, пов'язаних з ними технологіях та вебдизайнерських рішеннях, у загальному функціонуванні Інтернету виник цілий ряд нових явищ, що дозволив говорити про мережу другого покоління, або Web 2.0. Поняття "Web 2.0" є умовним терміном на позначення цих нових тенденцій, навіть певного повороту у загальній концепції Інтернету. Зміни полягають, зокрема, у посиленні комунікативності, співробітництва, безпечного використання інформації та загальному розвитку функціональності мережі. Термін вперше було вжито Тімом О'Рейлі у 2004 році на присвяченій інформаційним технологіям науковій конференції[Graham 2005]. Хоч англійська назва терміна вказує немов на нову версію мережі World Wide Web – 2.0 – насправді термін не означає оновлення технічних специфікацій, а лише зміни у шляхах використання мережі розробниками програмного забезпечення та кінцевими користувачами.

Слід також зазначити, що не всі дослідники вважають вживання цього терміна виправданим [Critical Perspectives on Web 2.0 2008]. Існує точка зору, що зміни були поступовими й еволюційними, і мережа від початку їх передбачала, тому говорити про різкий перелом не можна [Scholz 2008]. Втім, ми дотримуємось першої точки зору, тобто вважаємо зміни достатньо суттєвими, щоб говорити про мережу другого покоління.

Охопити все розмаїття сучасних напрямків розвитку мережевих технологій вкрай важко, тому ми торкнемось лише найсуттєвіших моментів.

Основною зміною у сприйнятті Інтернету з боку розробників сайтів та програмного забезпечення стало розуміння мережі як платформи [Hinchcliffe 2006]. У су-

часній мережі найбільших успіхів досягають і найбільш активно розвиваються саме ті проекти, які розглядають Інтернет як платформу, для якої створюються додатки-сайти. Ця загальна зміна нині обов'язково повинна враховуватися у прикладних лінгвістичних програмах.

Другою зміною у стратегії компаній-розробників стало розуміння того, що Інтернет створює передовсім можливість для надання послуг та сервісів, а не для продажу програмного забезпечення, що встановлюється на комп'ютері користувача. Саме в цій сфері зосереджена найбільша активність користувачів, а купівля чи вільне завантаження локального програмного забезпечення знаходиться на периферії. Компанія Google зробила ставку на надання послуг користувачам (пошук інформації тощо), тоді як компанія Netscape намагалась заробляти на продажі програмного забезпечення. Стрімкий успіх першої свідчить про правильно обрану стратегію напередодні домінування концепції мережі другого покоління [O'Reilly 2005]. Ця тенденція також повинна враховуватись у дослідженнях і практичних доробках з комп'ютерної лінгвістики. Також вона активно сприяє зростанню ролі лінгвістичного компоненту в Інтернет-проектах.

Сервіси передбачають обслуговування користувачів з усього світу, які є носіями різних мов, відповідно зростає необхідність оперативного перекладу інтерфейсів на якомога більшу кількість мов. Іноді це також торкається й вмісту (контенту) сайту. Розвиток світової культури показує, що тотальне домінування однієї мови поки що не справджується у повній мірі. Усі найбільш популярні та поширені Інтернет-сервіси не відмовляються від багатомовності. Сайти, що обмежуються однією мовою, як правило, неконкурентоспроможні, навіть якщо ця мова англійська.

Кожен сайт, фактично, є окремою програмою. Його незалежність від користувача знімає необхідність регулярних офіційних релізів та оновлень, розробники можуть оновлювати програму прямо в процесі роботи. Це спричиняє появу концепції "вічна бета", тобто вічної бета-версії програми, яка буде постійно оновлюватися і вдосконалюватися [O'Reilly 2005]. Паралельно розвивається сфера безкоштовного програмного забезпечення з відкритим кодом, який надає можливість людям з мінімальними знаннями програмування (а то й взагалі без них) створювати Інтернет-сайти і портали, що цілковито відповідають вимогам мережі другого покоління

Постійне оновлення програм передбачає оновлення і їхніх лінгвістичних складових. З урахуванням багатомовності більшості проектів це є досить складною і важливою проблемою. Відділення лінгвістичної інформації (наприклад, інтерфейсу) від безпосередніх програмних файлів сталося ще на попередньому етапі розвитку мережевих технологій, але тоді не можна було передбачити, настільки оперативно треба буде оновлювати цю інформацію. Ця потреба сприяє активному розвитку програмних додатків, які дозволяють редагувати інтерфейс у режимі реального часу редакторами проекту чи уповноваженими користувачами без будь-якого знання програмування і вебдизайну. Зростає потреба в електронних словниках, системах автоматичної перевірки орфографії та граматики, автоматичного перекладу, що часто вмонтовуються прямо в адміністраторську чи редакторську частину Інтернет-проектів. Тобто створюються своєрідні віртуальні робочі місця для редакторів і перекладачів, що у реальному часі підтримують проект.

Відбувається багато дрібних змін у програмному забезпеченні веб-сервісів. З'являються нові, більш уважні підходи до представлення результатів пошуку по мережі для користувачів. Розширюються можливості пошуку, який усе більше спирається на семантику. Створюються різноманітні лінгвістичні фільтри, що відсівають випадкові сторінки, ведуть пошук за синонімами тощо. Поштові служби починають використовувати фільтрування спаму та перевірку наявності у листах вірусів тощо. Системи фільтрації спаму, тобто небажаної реклами чи просто хуліганства, є прикладними лінгвістичними програмами. Ці програми автоматично аналізують потік електронних листів, що проходять крізь поштові сервери, відшукують часто повторювані тексти, або тексти, що містять певні ключові слова чи вже внесені в реєстр спаму і блокують доступ цих листів до користувачів поштових сервісів. Нині жодна поштова система не обходиться без подібних додатків, оскільки обсяг автоматично генерованого спаму складає абсолютну більшість електронних листів.

На зміну жорсткій категоризації змісту сайту приходить метод ключових слів, які додаються автором тексту у довільному порядку і потім відображаються за частотою вживання. Дуже активно це використовується у блогах, а також на сайтах, де публікуються зображення та інші медіа. Групування вмісту сайту, таким чином, є суттєвим поворотом в організації вмісту великих сайтів та й усієї мережі за семантичним принципом, структуризації матеріалів за змістом.

Дуже активно починають розвиватися бази даних. Нині вже майже не залишається сайтів, які б працювали без їх використання. Паралельно відбувається цікавий процес, коли компанія чи особа, яка володіє певними унікальними даними, може продавати право на їх використання різним сайтам [O'Reilly 2005]. Цю модель використовують і лінгвісти. Дослідники, що створюють певну базу даних (термінологічні чи орфографічні словники, парадигматичні бази даних, бази даних для систем автоматичного перекладу, корпуси текстів тощо) можуть надавати її за оплату компаніям-розробникам програмного забезпечення, які використовують її у тих чи інших цілях.

Розвивається співпраця між мережевими проектами і технічна можливість їхньої взаємодії. Виникають такі технології, як RSS, що дозволяють отримувати дані з сайту без його відвідування, відображати ці дані на інших сайтах чи в програмі на локальному комп'ютері. Взаємоінтеграція проектів теж вимагає прикладних лінгвістичних програм. Це, зокрема, програми семантичного аналізу, наприклад у сфері автоматичного чи автоматизованого підбору і групування новин, що збираються з різних Інтернет-ресурсів.

У мережі другого покоління радикально змінюється ставлення до користувача. Якщо спочатку він був просто читачем чи глядачем, то тепер він стає активним учасником сайтів [Graham 2005]. Популярність і (або) комерційна успішність прямо залежить від кількості залучених до проекту користувачів. Це може бути і залучення в якості автора текстів (Вікіпедія, Живий журнал) або інших медіа (YouTube, Flickr, Photo.net), автора коментарів чи відгуків на товар (Amazon), продавця чи покупця (eBay), людини, що доповнює базу даних проекту у тій чи інший сфері знань (Open Library), надає або завантажує певні дані (BitTorrent) чи навіть бере участь у розробці програмного забезпечення, на якому працює проект (Вікіпедія, та інші ресурси з відкритим кодом) тощо. Як вже згадувалось, увага до користувача авто-

матично тягне за собою зростання ваги лінгвістичного наповнення на сайті. Участь користувача у проекті передбачає надання йому певного лінгвістичного інструментарію – програми для редагування тексту, що часто включає автоматичну перевірку орфографії, термінологічні та перекладні словники відповідної галузі знань тощо.

Для залучення користувачів застосовуються різні способи зацікавлення, від різноманітних психологічних і психолінгвістичних прийомів до дуже помітної останнім часом тенденції до максимального спрощення інтерфейсу, його зручності [MacManus, Porter 2005], до відсутності явної реклами, прискорення швидкості завантаження сторінок (технологія AJAX). Паралельно розвиваються технології, що дозволяють перетворити сайт на аналог майже будь-якої звичайної програми, з велетенським набором функцій і зручною можливістю ними користуватися. Такими є нові поштові системи (Gmail), веб-редактори текстів, зображень тощо. Цілком можливо, що весь пакет офісних програм невдовзі буде доступний для користування просто через Інтернет, без потреби їхнього встановлення на локальному комп'ютері. Це стосується і таких традиційних лінгвістичних програм, як електронні словники та програми автоматичного перекладу тексту. Раніше вони поширювались для встановлення на комп'ютерах користувача. Нині вони доступні для користування через мережу з усіма перевагами нового "сервісного" підходу – регулярне оновлення баз даних і зменшення ціни чи навіть безкоштовність. У випадку автоматичних перекладачів у безкоштовній версії вводиться лише обмеження на кількість перекладеного тексту та присутня реклама.

У ставленні до користувача цікавим є підхід проекту онлайн-енциклопедії "Вікіпедія". Це так звана концепція радикальної довіри, коли кожен користувач може змінювати вміст сторінок сайту. Розпочиналось усе як експеримент, але успіх був неймовірний. Цей повністю некомерційний проект нині входить у 100 найвідвідуваніших сайтів мережі Інтернет, і залишив далеко позаду усіх конкурентів у своїй сфері за кількістю енциклопедичних статей – в англійському розділі сайту їх більше 2,5 мільйонів. Цей проект, зрозуміло, не є строго науковим, але за охопленим матеріалом він недосяжний. У проекті цікаво реалізовано семантичне структурування матеріалів та їхній багатомовний характер (тобто кожна стаття має відсилання на цю ж статтю іншими мовами, що представлені в окремих мовних версіях Вікіпедії). Хоч структурування і багатомовність спочатку робились вручну користувачами, останнім часом починають здійснюватись спроби часткової автоматизації цих процесів.

Окрім залучення користувачів до участі в проекті, важливу роль відіграє організація їх співпраці, своєрідного колективного інтелекту. Користувачі, що здобули довіру, можуть ставати модераторами або адміністраторами проекту, що вибудовує певну ієрархію і дозволяє краще організувати загальну роботу. Для організації роботи такого віртуального колективу нерідко використовуються напрацювання психолінгвістики та інших суміжних з лінгвістикою дисциплін.

Розвиток електронних бібліотек гальмується законодавством, що забороняє вільне розповсюдження захищених авторським правом текстів. Утім, це не завадило виникненню таких потужних проектів мережі другого покоління, як Google Books та Open Library. Вони містять мільйони повнотекстових (відсканованих) книг, які додають самі користувачі. Тут теж використовуються системи автоматичної та автоматизова-

ної обробки текстової інформації, системи автоматичного розпізнавання текстів із відсканованих зображень, різноманітні бібліографічні системи опису і каталогізації.

Активно розвиваються системи дистанційного навчання, які теж часто є прикладними лінгвістичними програмами чи принаймні пов'язані з лінгвістикою. У тому числі активізується дистанційне навчання природним мовам (Mova.info).

Важливим етапом у становленні мережі другого покоління стала поява та активний розвиток блогів. Це сервіси, що прийшли на зміну персональним веб-сторінкам і дозволяли користувачам додавати на власну сторінку довільні записи у хронологічний послідовності. Їх також називають електронними щоденниками або журналами. Привабливості цьому явищу додав його комунікативний аспект, чого не було на персональних сторінках. Користувачі могли не просто додавати свої записи на власній сторінці, а й коментувати записи інших на їхніх сторінках, знайомитися, спілкуватися тощо. Блоги призвели до геометричного зростання різноманітної текстової інформації у мережі Інтернет і, відповідно, до потреби у нових принципах її автоматичної обробки пошуковими системами, а також до розвитку програм редагування тексту через веб-оглядач.

Розвиток блогів вилився у наступний прорив у мережевих технологіях та концепціях – появу соціальних мереж. У таких проектах зареєстровані користувачі створюють закриту для сторонніх і захищену (як декларується) спільноту, де вони подають правдиві дані про себе і за такими ж даними можуть знаходити своїх колишніх друзів, однокласників, однокурсників, родичів, чи нових знайомих за професійними інтересами чи іншими зацікавленнями [Vickery, Wunsch-Vincent 2007]. На сьогоднішній момент це один з найбільш активно зростаючих сегментів Інтернету на пострадянському просторі (ВКонтакте, Одноклассники) та й у світі теж (Facebook). Зовсім недавно з'явились проекти, які дозволяють користувачу в одному місці тримати всі свої реєстрації на будь-яких інших сайтах, редагувати їх, публікувати на інших сайтах повідомлення і переглядати їхній вміст тощо. Створюється своєрідне віртуальне "робоче місце" користувача, яке дозволить йому мати всі потрібні йому сайти в одному місці і не виникатиме потреба постійно безпосередньо їх відвідувати. Ці проекти та соціальні мережі є дуже складними комплексними програмами, що включають у себе багато компонентів, зокрема й лінгвістичних, які ми розглянули вище.

Навіть при побіжному огляді особливостей мережі другого покоління, очевидною є значна роль практично всюди комп'ютерної лінгвістики. Зростає комунікативний аспект мережі, а оскільки комунікація напряму пов'язана з мовою, то зростає і вага досліджень та прикладних розробок з комп'ютерної лінгвістики у функціонуванні мережі. Комп'ютерна лінгвістика відіграє суттєву, якщо не ключову, роль на новому етапі розвитку Інтернету, тому дуже важливим є врахування усіх нових мережевих тенденцій як у майбутніх теоретичних дослідженнях, так і в прикладних напрацюваннях.

Теоретические исследования в сфере компьютерной лингвистики и их прикладной характер являются обязательными элементами большинства Интернет-проектов. В статье рассматривается роль прикладной лингвистической составляющей в принципиально новых сетевых концептуальных и технологических явлениях, которые условно называются сетью второго поколения или Web 2.0.

Ключевые слова: *компьютерная лингвистика, прикладная лингвистика, Web 2.0, Интернет, информационные технологии, обработка природного языка.*

Theoretical researches in the field of computational linguistics of natural language processing and its practical implementations are usual elements of most of modern Internet-projects. Article analyses the role of applied linguistic components in new conceptual and technological phenomena that is usually refered to as Web 2.0.

Key words: *computational linguistics, applied linguistics, natural language processing, Web 2.0, Internet, IT.*

Література:

1. *Critical Perspectives of Web 2.0.* Special issue of First Monday. – Vol. 13. – № 3. – 2008 // Електронний ресурс: http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/issue/view/263/showToc (16.03.2009).

2. *Graham P.* Web 2.0 / P. Graham // Електронний ресурс: http://www.paulgraham.com/web20.html.

3. *Hinchcliffe D.* The State of Web 2.0. / D. Hinchcliffe // Електронний ресурс: http://web2.wsj2.com/the_state_of_web_20.htm (16.03.2009).

4. *MacManus R., Porter J.* Web 2.0 for Designers / R. MacManus, J. Porter // Digital Web Magazine // Електронний ресурс: http://www.digital-web.com/articles/web_2_for_designers.

5. *O'Reilly T.* What Is Web 2.0 / T. O'Reilly // Електронний ресурс: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

6. *Scholz T.* Market Ideology and the Myths of Web 2.0 / T. Scholz // First Monday. – Vol. 13. – № 3. – 2008 // Електронний ресурс: http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2138/1945.

7. *Singel R.* Are You Ready for Web 2.0? / R. Singel // Електронний ресурс: http://www.wired.com/news/technology/0,1282,69114,00.html.

8. *Vickery G.* Participative Web and User-Created Content: Web 2.0, Wikis and Social Networking // OECD/ G. Vickery, S. Wunsch-Vincent // Електроннийресурс: http://www.oecd.org/document/40/0,3343,en_2649_201185_39428648_1_1_1_1,00.html.

9. *Дарчук Н.П.* Комп'ютерна лінгвістика (автоматичне опрацювання тексту) / Н.П. Дарчук. – К.: Видавничо-поліграфічний центр "Київський університет", 2008. – 351 с.

10. *Дерба С.М.* Словник з української термінології прикладної (комп'ютерної) лінгвістики / С.М. Дерба. – К.: 2007. – 325 с.