УДК 004.522

ІНТЕРНЕТ-РЕСУРСИ І КОМП'ЮТЕРНИЙ ПЕРЕКЛАД

Франчук Наталія Петрівна,

доцент кафедри теоретичних основ інформатики Інституту інформатики НПУ імені М. П. Драгоманова, м. Київ.

Анотація. У статті йдеться про комп'ютерний переклад Web-сторінок різними мовами. Розглядаються питання компанії Google стосовно стратегії діяльності напередодні домінування в мережі Інтернет концепції мережі другого покоління. Розглядається виконання перекладу тексту за допомогою інструментів Google Chrome.

Ключові слова: Web-сторінка, глобальна мережа Інтернет, технології RSS, комп'ютерний переклад, соціальні мережі.

На початку нового тисячоліття вченими було розроблено низку нових сервісів, пов'язаних з мережевими Інтернет-технологіями і веб-дизайнерськими рішеннями. Практичне масове втілення цих сервісів дало підстави говорити про мережу другого покоління або **Web 2.0.** Саме поняття «Web 2.0» є умовним терміном, що вказує на зміни концепції використання глобальної мережі Інтернет. Зміни полягають, зокрема, у посиленні в цих сервісах функцій комунікативності, співробітництва, безпечного використання даних користувачами.

Основні зміни у стратегії компаній-розробників відбулися після усвідомлення того, що за допомогою сервісних функцій глобальної мережі Інтернет забезпечуються передусім умови для постійного надання користувацьких сервісів, а не лише для разового продажу програмного забезпечення, що потрібно інсталювати на комп'ютері користувача.

Саме в постійному отриманні сервісних функцій зосереджена нині найбільша активність користувачів, а послуги купівлі чи вільного завантаження користувачами локального програмного забезпечення поступово стають неактуальними. Так, наприклад, компанія Google зробила ставку на надання постійних послуг користувачам (пошук даних, користування електронною поштою, календарем, картою, перекладачами, спільними документами тощо), тоді як компанія Netscape намагалась заробляти переважно на продажах програмного забезпечення. Стрімкий фінансовий успіх першої компанії свідчить про правильно обрану стратегію діяльності напередодні домінування в мережі Інтернет концепції мережі другого покоління [2].

На цю тенденцію також слід зважати у проведенні досліджень і створенні практичних розробок, зокрема й у використанні інтернет-сервісів для комп'ютерного перекладу текстів.

Переклад текстів — одна з перших функцій, яку людина спробувала виконати за допомогою комп'ютера. Усього через кілька років після створення перших ЕОМ було розроблено й програми для машинного перекладу. Датою народження машинного перекладу (*machine translation*) як галузі досліджень прийнято вважати 1947 р. [6].

До епохи масового поширення персональних комп'ютерів машинний переклад міг бути лише цікавим об'єктом наукових досліджень, ніж важливою масовою сферою застосування обчислювальної техніки.

Причинами цього були:

- **висока вартість машинного часу роботи ЕОМ**, з огляду на той факт, що кожну обчислювальну машину обслуговувала велика група фахівців (системних програмістів, інженерів, техніків і операторів), для кожної машини було потрібне окреме велике, спеціально обладнане приміщення, окреме спеціально створене (саме для конкретної ЕОМ) програмне забезпечення і т. п., «комп'ютерний час» був дуже і дуже дорогим);
- **колективне використання ресурсів комп'ютера.** Саме це не давало змоги конкретному користувачеві негайно звернутися до електронного помічника, що зводило нанівець найважливішу перевагу машинного перекладу перед звичайним — його оперативність.

Саме масове поширення персональних комп'ютерів (ПК) стало ефективним додатковим стимулом для вдосконалювання комп'ютерного перекладу (особливо після створення комп'ютерів Apple II у 1977 р. і комп'ютерів IBM PC у 1981 р.). Поновленню досліджень з комп'ютерного перекладу сприяли також пришвидшення темпів розвитку як обчислювальної техніки, так і програмного забезпечення. Так, у 70-ті роки XX ст. набула поширення система автоматизованого перекладу **SYSTRAN**. У 1974–75 роки система була використана аерокосмічною асоціацією NASA для перекладу документів проекту «Союз-Аполлон». До кінця 80-х років XX ст. за допомогою саме цієї системи виконували переклад документів з кількох мов щорічно обсягом вже близько 100 000 сторінок. Розвитку комп'ютерного перекладу сприяло ще й зростання інтересу дослідників і проектувальників до проблем штучного інтелекту (тут явно переважали лінгвістичні аспекти) і до комп'ютерного пошуку даних [3].

Починаючи з 80-х років XX ст., коли вартість машинного часу помітно знизилась, а доступ до комп'ютера можна було користувачеві одержати в будь-який час, виконання машинного перекладу документів стало економічно вигідним. У ці й наступні роки завдяки удосконаленню програм можна було досить точно перекладати багато видів текстів. 90-ті роки XX ст. можна вважати справжньою «епохою Відродження» у сфері розвитку комп'ютерного перекладу, що пов'язано не тільки з широкими можливостями використання ПК і доступністю для користувачів нових технічних засобів (зокрема сканерів), але і з застосуванням комп'ютерних мереж, зокрема глобальної мережі Інтернет.

Наприклад, створення Європейської Інформаційної Мережі (EURONET DIANA) стимулювало діяльність фахівців зі створення систем автоматизованого перекладу. У 1982 р. було оголошено про створення європейської програми EUROTRA, метою реалізації якої була розробка системи комп'ютерного перекладу для всіх європейських мов. Спочатку проект оцінювався в 12 млн. доларів США, але вже в 1987 р. фахівці визначили сумарні витрати на цей проект понад 160 млн. доларів [1].

Використання глобальної мережі Інтернет об'єднало мільйони людей, що говорять різними мовами, у єдиний інформаційний простір. За допомогою сервісів і служб глобальної мережі Інтернет обслуговуються користувачі з усього світу, які є носіями різних мов, отже, зростає необхідність оперативного перекладу інтерфейсів програм на якомога більшу кількість мов. У більшості випадків це також стосується й вмісту (контенту) сайту. Розвиток світової культури показує, що тотальне домінування однієї мови поки що є неможливим і недоцільним. У всіх найпопулярніших і поширених Інтернет-сервісах і службах є функції забезпечення багатомовності. Сайти, на яких контент подано лише однією мовою, як правило, неконкурентоспроможні. Нині у світі є тенденція до домінування англійської мови, але:

- є користувачі, які англійською зовсім не володіють чи засвоїли її дуже слабко;
- велика кількість Web-сторінок, інтерфейс і контент яких подано не англійською мовою.

Для полегшення перегляду користувачем Web-сторінок, описаних незнайомою користувачеві мовою, фахівцями було розроблено спеціальні додатки до браузерів, за допомогою яких можна виконувати переклад обраних користувачем фрагментів Web-сторінки або всієї Web-сторінки, що переглядається. Для цього користувачеві необхідно лише скопіювати частину тексту і вставити її у відповідне поле програми або «натиснути» на спеціальну кнопку меню.

Прикладом такого комп'ютеризованого перекладача є **Google Translate** — це сервісний продукт від компанії Google, за допомогою якого можна автоматично перекладати слова, фрази та web-сторінки з однієї мови на іншу. У пошуковій системі Google використовується програмне забезпечення, розроблене фахівцями компанії Google, що призначене для перекладу на основі *статистичного машинного перекладу*. Користувач вставляє у вікно програми текст, поданий мовою оригіналу, і вказує мову, якою цей текст потрібно подати. З вересня 2008 р. програма має функції підтримки й перекладу українською мовою.

Нині над проблемами машинного перекладу працюють фахівці відомих компаній, таких як: SYSTRAN Software Inc., Logos Corp., Globalink Inc., Alis Technologies Inc., Toshiba Corp., Compu Serve, Fujitsu Corp., TRADOS Inc., Промт та інші. Є компанії, що спеціалізуються лише в галузі машинного перекладу, зокрема компанія SAP AG, яка є європейським лідером у розробці програмного забезпечення і протягом багатьох років використовує системи машинного перекладу різних виробників під час локалізації своїх програмних продуктів. Утворено і службу машинного перекладу при комісії Європейського Союзу (обсяг перекладу документів у комісії перевищує 2,5 млн. сторінок щорічно; переклади всіх документів виконуються оперативно 11-тьма офіційними мовами, забезпечують таку роботу 1100 перекладачів, 100 лінгвістів, 100 менеджерів і 500 секретарів) [4].

Постійне оновлення програм для перекладу передбачає оновлення і їх мовних складових. З урахуванням багатомовності більшості проектів це є досить складною і важливою проблемою. Відокремлення файлів, що містять дані про інтерфейс програми (описаний різними мовами), від файлів із програмним кодом сталося ще на попередньому етапі розвитку мережевих технологій, але тоді не можна було передбачити, наскільки оперативно потрібно буде оновлювати ці дані. Розв'язання таких проблем сприяє активному розвитку програмних засобів, за допомогою використання яких можна редагувати інтерфейс у режимі реального часу редакторами проекту (чи уповноваженими користувачами) без будь-яких знань з програмування і з веб-дизайну.

Для надання якісніших й оперативніших сервісів користувацької багатомовності в Інтернет-просторі зростає потреба в електронних словниках, системах автоматичної перевірки орфографії і граматики, автоматичного перекладу, що часто вмонтовуються безпосередньо в адміністративну чи редакторську частину Інтернет-проектів (сайтів чи порталів). Тобто створюються своєрідні віртуальні робочі місця для редакторів і перекладачів, які у реальному часі підтримують проект.

Відбуваються також зміни у програмному забезпеченні веб-сервісів. Для користувачів розробляються нові, більш виважені підходи до подання результатів пошуку в мережі Інтернет. Вдосконалюються функції для виконання пошуку, який все більше спирається на семантику. Розробники пошукових програм пропонують різноманітні мовні фільтри, за допомогою яких «відсіваються випадкові сторінки», ведеться пошук за синонімами тощо.

У поштових службах починають використовувати функції фільтрування спаму й перевірки наявності у листах вірусів тощо. Системи фільтрації спаму, тобто небажаної реклами чи непристойних повідомлень, базуються на прикладних мовних програмах. За допомогою таких програм автоматично аналізується потік електронних листів, що проходить крізь поштові сервери, в електронних листах відшукуються часто повторювані тексти (різними мовами), або тексти, що містять певні ключові слова, які вже внесені в реєстр спаму, і якщо такі фрагменти виявлено, то до цих листів блокується доступ користувачів поштових сервісів. На сьогодні жодна поштова система не використовується без подібних програмних засобів, оскільки обсяг автоматично генерованого спаму нині складає абсолютну більшість у потоках електронних листів.

На зміну жорсткій категоризації вмісту сайту приходить метод ключових слів, які додаються автором тексту у довільному порядку і потім відображаються за частотою використання. Це переважно використовується у блогах (*блог* — це веб-сайт, головний зміст якого — записи, зображення чи мультимедіа, що регулярно додаються користувачем, для блогів характерні короткі записи тимчасової значущості), а також на сайтах, де публікуються різноманітні зображення та інші медіадані. Групування вмісту сайту, отже, є суттєвим поворотом в організації контенту на великих сайтах й у всій мережі Інтернет за семантичним принципом, структуризації матеріалів за вмістом.

Розробники сайтів дуже активно починають використовувати *бази даних*. На сьогодні вже майже не залишається сайтів, які б функціонували без їх використання. Паралельно відбувається цікавий процес, коли компанія чи особа, яка володіє певними унікальними даними, може продавати право на їх використання різним сайтам. Це використовують і лінгвісти-дослідники, які створюють певну базу даних (термінологічні чи орфографічні словники, парадигматичні бази даних, бази даних для систем автоматичного перекладу тощо) і можуть надавати її за оплату компаніям-розробникам програмного забезпечення, які використовують такі бази даних з певною метою.

Слід також звернути увагу й на технології RSS (*Really Simple Syndication* — просте отримання даних). RSS — це система XML-форматів, що використовується для публікації й передавання даних, що часто змінюються, наприклад нових записів у блозі, заголовків новин, анонсів статей, зображень, аудіо- і відеоматеріалів (рис. 1).

За допомогою RSS можна отримувати дані із сайту, не відвідуючи сам сайт, відображати ці дані на інших сайтах чи в спеціальній програмі на локальному комп'ютері. Взаємоінтеграція проектів неможлива без використання прикладних лінгвістичних програм. Це, зокрема, програми семантичного аналізу, наприклад у сфері автоматичного чи автоматизованого добору і групування новин, що збираються з різних Інтернет-ресурсів [5].

Для залучення користувачів до цілеспрямованого відвідування конкретних Інтернет-ресурсів розробники застосовують різні способи мотивації користувачів: від різноманітних психологічних і психолінгвістичних прийомів до дуже помітної останнім часом тенденції максимального спрощення інтерфейсу, його зручності, прискорення швидкості завантаження сторінок (технологія AJAX — (*Asynchronous JavaScript And XML*) — підхід до побудови користувацьких інтерфейсів веб-застосувань, за яким відправлення запитів на сервер і завантаження потрібних користувачеві даних відбуваються у фоновому режимі без перезавантаження веб-сторінки).

Такі саме тенденції характерні й для таких традиційних лінгвістичних програм, як електронні словники і програми автоматичного перекладу текстів. Раніше такі програмні продукти поширювали лише для інсталяції на комп'ютерах користувача. Нині користувачеві не потрібно їх інсталювати на своєму комп'ютері, адже ці програмні продукти доступні для використання через мережу Інтернет з усіма перевагами нового «сервісного» підходу — регулярне оновлення баз даних і зменшення їхньої ціни (чи, навіть, безкоштовність). Нині у разі використання автоматичних перекладачів (наприклад, **Google Translate** — служба безкоштовного онлайн-перекладу пошукової системи Google, за допомогою якої можна миттєво перекладати тексти і веб-сторінки) у безкоштовній версії вводиться лише обмеження на кількість перекладеного тексту і наявна реклама.

Слід звернути увагу на концепцію так званої он-лайн-енциклопедії «Вікіпедія» — це концепція довіри, коли будь-який користувач може змінювати вміст сторінок цього сайту. Розпочиналось усе як експеримент, але успіх був неймовірний. Цей повністю некомерційний

Рис. 1

проект нині входить до першої сотні найвідвідуваніших сайтів глобальної мережі Інтернет, далеко позаду залишилися всі конкуренти у цій сфері: у рейтингу за кількістю енциклопедичних статей — в англійському розділі сайту їх більше 2,5 мільйонів [8]. Цей проект, зрозуміло, не може бути суто науковим, але за широтою охопленої тематики і за кількістю матеріалу — він недосяжний. У проекті цікаво реалізовано семантичне структурування матеріалів та їх багатомовний характер (майже кожна стаття містить посилання на цю саму статтю іншими мовами, що подані в окремих мовних версіях Вікіпедії). Хоч структурування і багатомовність матеріалів спочатку робились «вручну» користувачами, останнім часом у цьому проекті спостерігаються спроби часткової автоматизації цих процесів. Окрім залучення користувачів до участі в створенні матеріалів для проекту важливу роль відіграє організація їхньої взаємодії, адже тут можна спостерігати утворення своєрідного «колективного інтелекту». Користувачі, які здобули довіру (мають не менше 1000 редагувань в основному просторі, 3 місяці стажу та 10 редагувань службового простору), можуть ставати модераторами або навіть адміністраторами проекту, що вибудовує певну ієрархію, і дозволяє краще організувати спільну роботу (рис. 2).

Нині в Інтернет-просторі розвиток електронних бібліотек гальмується законодавством про авторські права, що забороняє вільне розповсюдження захищених авторським правом текстів, аудіо та відео. Але це не стало перешкодою у створенні таких потужних проектів мереж другого покоління, як проект **Google Books** (сервіс від компанії Google, за допомогою якого можна здійснювати повнотекстовий пошук усередині книжок і журналів, які компанія Google сканує і розміщує у своїй базі даних відсканованих книжок і журналів) та проект **Open Library**. У них містяться мільйони повнотекстових (відсканованих) книжок, які додаються самими користувачами Мережі. Тут теж активно використовуються системи автоматичного й автоматизованого опрацювання текстових даних, розпізнавання поданих різними мовами текстів із відсканованих зображень, різноманітні бібліографічні системи опису і каталогізації.

Стрімко розвиваються й системи дистанційного навчання, у яких також часто використовуються прикладні лінгвістичні програми

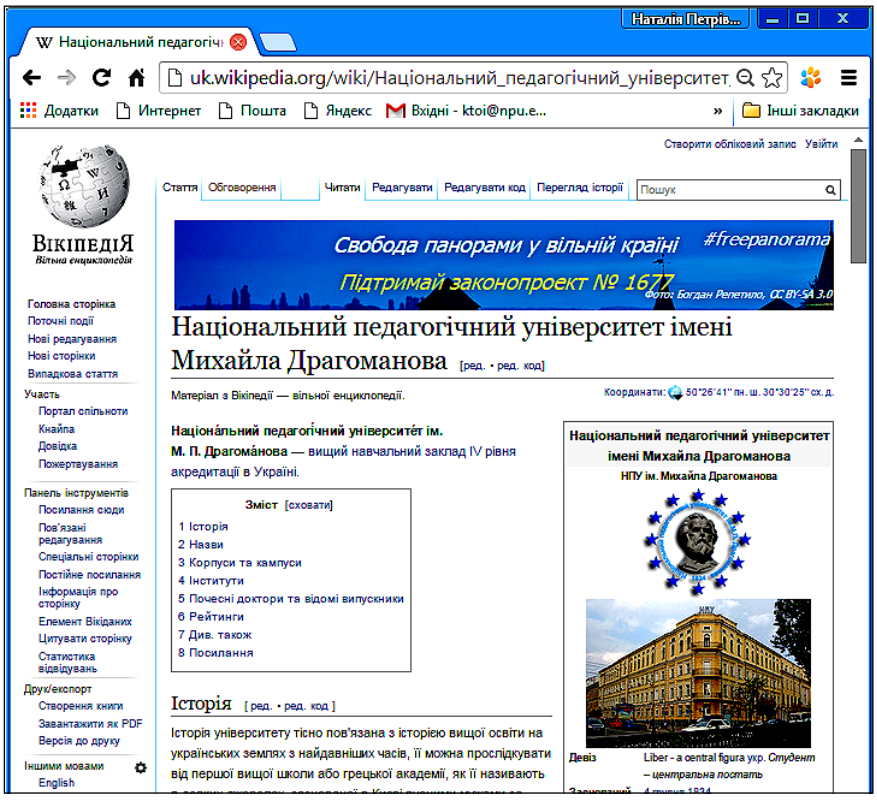

Рис. 2

чи, принаймні, пов'язані з лінгвістикою. У тому числі активізується дистанційне навчання природних мов. Наприклад, за допомогою порталу Mova.info — (http://www.mova.info/) можна здійснювати систематичну довідково-пошукову роботу стосовно української мови й української лінгвістики в глобальній мережі Інтернет.

Важливим етапом у становленні Мережі другого покоління стала поява і стрімкий розвиток так званих *блогів*. Це сервіси, що прийшли на зміну персональним веб-сторінкам, за допомогою яких користувачі додають на власну сторінку довільні записи у хронологічній послідовності. Їх також називають *електронними щоденниками* або *журналами*. Популярності цьому сервісу додав його комунікативний аспект, чого не було на персональних сторінках (сайтах). Користувачі можуть додавати записи на власній сторінці й коментувати записи інших користувачів на їхніх сторінках, знайомитися, спілкуватися тощо.

Масове використання блогів призвело до зростання в геометричній прогресії кількості різноманітних текстових даних у глобальній мережі Інтернет і, відповідно, до потреби в нових принципах автоматичного опрацювання таких даних за допомогою пошукових систем, а також до розширення функцій програм для редагування текстів через веб-браузер.

Активне масове використання блогів спонукало до розвитку таких мережевих технологій і концепцій — *соціальних мереж*.

У таких проектах зареєстровані користувачі мають змогу створювати закриту для сторонніх і захищену (як декларується) спільноту, учасники якої 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 вподобаннями [1].

Нині це один з найбільш стрімко зростаючих сегментів глобальної мережі Інтернет на пострадянському просторі (ВКонтакте — vk.com, Одноклассники — www.odnoklassniki.ua) та й у всьому світі теж (Facebook — www.facebook.com). Нещодавно було розроблено проекти, за допомогою яких користувач в одному місці може мати всі свої реєстраційні дані для будь-яких інших сайтів, крім цього користувач має змогу редагувати їх, публікувати на інших сайтах повідомлення і переглядати їхній вміст тощо. Отже, утворюється своєрідне *віртуальне «робоче середовище» користувача*, за допомогою якого він може мати всі потрібні йому дані з різних сайтів в одному місці. Отже, відпадає потреба постійно відвідувати багато сайтів (рис. 3).

З поширенням вище описаних функцій зростає комунікативний аспект Мережі, а оскільки комунікації між людьми прямо пов'язані з отриманням повідомлень, поданих різними мовами, то зростає і важливість досліджень та прикладних розробок з комп'ютерного, комп'ютеризованого та напівавтоматичного перекладу повідомлень, отриманих з Мережі. Використання функцій комп'ютерного перекладу відіграє суттєву (одну з ключових) роль на новому етапі розвитку глобальної мережі Інтернет, тому найактуальнішим нині є врахування всіх тенденцій розвитку мережевих технологій як у майбутніх теоретичних дослідженнях, так і в прикладних напрацюваннях.

Слід звернути увагу на інструменти перекладу веб-сторінок «на льоту», наприклад, на використання компонента веб-браузера **Google Chrome** для перекладу веб-сторінки потрібною мовою. За допомогою вбудованої панелі перекладу в Google Chrome користувач може читати більше веб-сторінок, незалежно від мови їх подання.

Розглянемо приклад виконання перекладу тексту за допомогою інструментів Google Chrome. Під час відкривання веб-сторінки, що містить контент мовою, якої немає серед пропонованих мов на цій вебсторінці, у Google Chrome відкривається панель перекладу вгорі сторінки. Увімкнення і вимкнення панелі перекладу в Google Chrome можна виконати так:

- відкрити головне меню на панелі інструментів веб-браузера Google Chrome;
- вибрати пункт меню «**Налаштування**»;
- у вибраному пункті перейти до пункту меню «**Показати розширені налаштування**».

Цю функцію можна налаштувати, поставивши чи знявши прапорець біля пункту «**Пропонувати перекладати сторінки, мова яких відмінна від тієї, якою я читаю**» в розділі «**Мови**».

Коли панель перекладу ввімкнена, можна налаштувати додаткові параметри. Для цього слід «натиснути» на кнопку «**Параметри**» панелі перекладу і вибрати вказані нижче налаштування, щоб управляти відображенням панелі перекладу для певних сайтів і мов:

- вибрати «**Завжди перекладати цю мовну пару: [мова] — украї-**

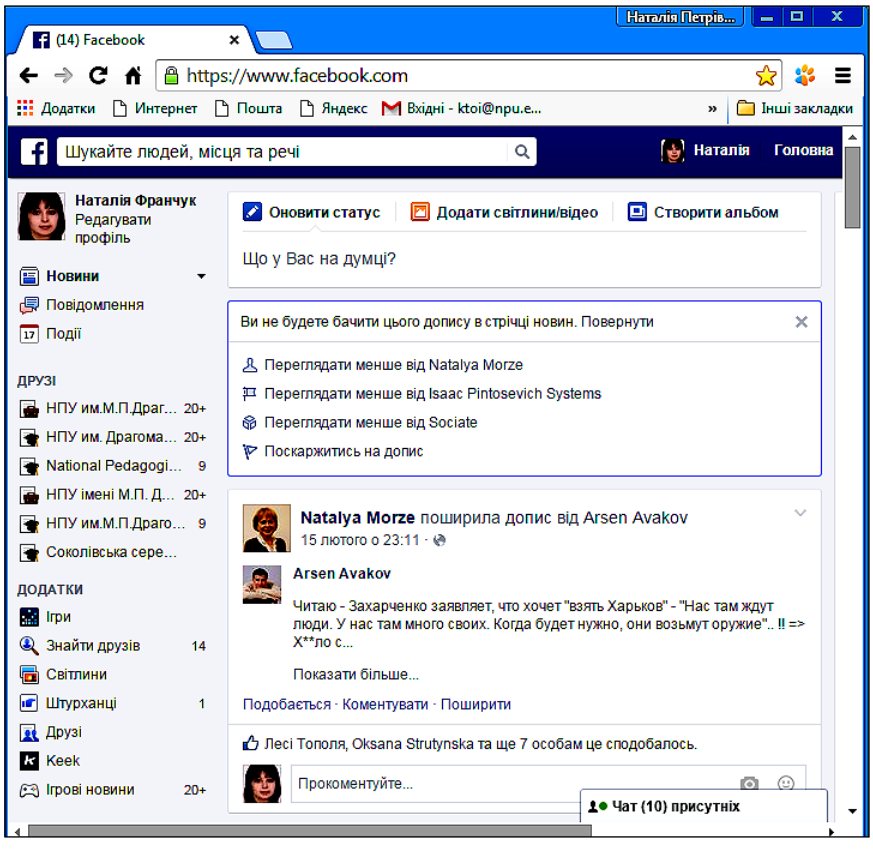

Рис. 3

нська», щоб автоматично здійснювався переклад сторінок, написаних саме цією мовою;

- вибрати **«Ніколи не перекладати з такої мови: [*мова*]»**, якщо не потрібно, щоб панель перекладу відображалася для цієї мови;
- вибрати **«Ніколи не перекладати цей сайт»**, якщо не потрібно, щоб панель перекладу відображалася для сторінок веб-сайта, який переглядається.

Нині машинний переклад далекий від ідеалу, але з використанням таких інструментів можна отримати певну допомогу під час роботи з ресурсами, які подані іншими мовами.

Дуже зручними у використанні є вбудовані панелі перекладу, за допомогою яких можна читати більше веб-сторінок, незалежно від того, якою мовою вони подані, з яких дізнатися про нові відомості. Під час відвідування сторінок, описаних мовою, відмінною від використовуваної в інтерфейсі програми для перегляду веб-сторінок, у верхній частині сторінки автоматично з'являється панель з пропозицією перекласти веб-сторінку. Користувач може вибрати мову веб-сторінки і сайти, зміст яких у майбутньому перекладати не потрібно, це буде здійснюватися автоматично під час завантаження сайту. Також користувач може повністю відключити функцію перекладу в налаштуваннях, вибравши в допоміжному вікні послугу **Параметри** (рис. 4).

Наразі на сьогодні найкращими поки що залишаються переклади спеціально підготовленими фахівцями, які досконало володіють мовами, предметними знаннями у відповідній галузі, а також сучасними інформаційно-комунікаційними технологіями. Широке залучення майбутніх учителів інформатики до розв'язування задач локалізації програмних засобів з відкритим кодом сприяє значному покращенню їхньої фахової підготовки.

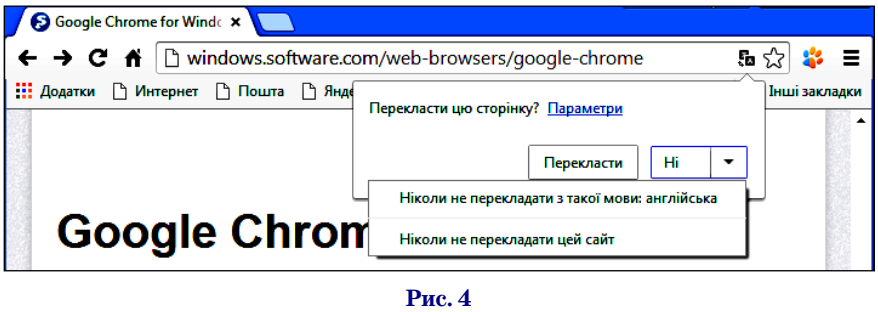

Рис. 4

* * *

Франчук Н. П. Интернет-ресурсы и компьютерный перевод

Аннотация. В статье говорится о компьютерном переводе Web-страниц разными языками. Рассматриваются вопросы компании Google относительно стратегии деятельности накануне доминирования в сети Интернет концепции сети второго поколения. Рассматривается выполнение перевода текста с помощью инструментов Google Chrome.

Ключевые слова: Web-страница, глобальная сеть Интернет, технологии RSS, компьютерный перевод, социальные сети.

* * *

Franchuk Natalia. Internet resources and computer translation

Summary. The article deals with computer translation of Web-pages in different languages. The questions in relation Google's strategy of domination on the eve of the Internet network concept of the second generation. Consider the translation of the text using the tools Google Chrome.

Keywords: Web-page, global network of Internet, technologies RSS, computer translation, social networks.

Література

1. *Анисимов А. В.* Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык. — Киев : Наук.думка, 1991. — 208 с.
2. Війна браузерів — Вікіпедія [Електронний ресурс]. — Режим доступу: http://uk.wikipedia.org/wiki/Війна_браузерів.
3. Історія машинного перекладу — Вікіпедія [Електронний ресурс] //Wikipedia — 2012. —Режим доступу до бібліотеки : http://uk.wikipedia.org/wiki/Історія_машинного_перекладу.
4. *Сущук О. А.* Міжнародні інформаційні системи: навчальний посібник. — К. : ІЗМН, 1999. — 224 с.
5. *Франчук Н. П.* Комп'ютеризований переклад з використанням web-орієнтованих програмних засобів // Науковий часопис НПУ імені М. П. Драгоманова. Серія №2. Комп'ютерно-орієнтовані системи навчання : зб. наук. праць / Редрада. — К. : НПУ імені М. П. Драгоманова, 2012. — № 13 (20). — С. 120–124.
6. *Франчук Н. П.* Стан та перспективи технологій машинного перекладу тексту // Теорія та методика електронного навчання : збірник наукових праць. Випуск ІІІ. — Кривий Ріг : Видавничий відділ НМетАУ, 2012. — С. 319–325.
7. Web 2.0 for development [Electronic resource] // Web 2.0. — Mode of access : http://en.wikipedia.org/wiki/ Web_2.0_for_development.
8. Wikipedia [Electronic resource] // Web 2.0. — Mode of access: http://en.wikipedia.org.