О.Й. Максимів

КОРПУС ПЕРСЬКОЇ МОВИ ЯК ДЖЕРЕЛО МАТЕРІАЛУ ДЛЯ ЧАСТОТНОГО СЛОВНИКА

Найновіші праці, присвячені як комп'ютерній лінгвістиці [див., напр.: Meyer 2004; Корпусна... 2005; Демська-Кульчицька$_1$ 2005; Прикладна... 2007 та ін.], так і укладанню частотних словників [див., напр.: Алексеев 2001; Bybee 2001; Перебийніс 2002; Мартинюк 2003; Бук 2006 та ін.], підтверджують, що ці обидві галузі мовознавства на сьогодні є актуальними й активно розвиваються. Проте усі відомі нам частотні словники укладені вручну чи напівавтоматично на окремо підібраному для цього матеріалі. Як правило, формується вибірка, обсяг якої автори словника вважають достатньо великим та репрезентативним для обраної мови чи субмови і достатньо малим, аби його опрацювати за певний осяжний проміжок часу [див., напр.: Андрющенко 1968; Частотный... 1977]. Звісно, що такий компроміс обмежує можливість залучення якомога більшого обсягу матеріалу, що є однією з основних вимог достовірності статистичного дослідження [див.: Засорина$_2$ 1966; Муравицька 1967; Перебийніс 1985; Алексеев 2001; Демська-Кульчицька$_1$ 2005]. З другого боку, не завжди при випадковому підборі навіть великої кількості авторів можна уникнути неточностей і навіть помилок; аналіз меншого за обсягом, але цілеспрямовано і свідомо підібраного матеріалу може дати дуже точні висновки [Троянская 1967; Арнольд 1991].

Мета цієї розвідки полягає в тому, щоб з'ясувати, чи можливо укласти частотний словник на основі корпусу певної мови і якою в такому разі буде процедура укладання такого словника.

За типом джерел частотні словники бувають або загальними, тобто укладеними на матеріалі змішаних текстів, які пропорційно відображають різні мовленнєві жанри, або галузевими, які базуються на текстах певного жанру чи галузі [див., напр.: Малаховский 1966; Андрющенко 1968]. Також існують частотні словники мови письменників або окремих творів. Матеріалом для укладання частотного словника може правити як індивідуальний текст, так і набір текстів, причому в обох випадках за основу беруть або цілісні (завершені) тексти, або уривки (вибірки) із текстів. На практиці у питаннях дозування і визначення обсягу великого частотного словника спостерігаються значні розбіжності [див. дет.: Тулдава 1987]. Одне з важливих питань, яке доводиться вирішувати авторові частотного словника, стосується розміру мінімальної вибірки, тобто довжини тих фрагментів тексту, з яких вручну складається вибірковий корпус [див.: Перебийніс 1985; Тулдава 1987; Алексеев 2001].

Різноманітність вибірки забезпечують, здебільшого виділяючи чотири-п'ять стандартних „жанрів" або функціональних стилів мовлення – драму, художню літературу, листування, періодичну літературу та „технічну" літературу [див., напр.: Андрющенко 1968; Алексеев 1971; Частотный... 1977]. У деяких словниках окремо не виділяють жанру листування, але залучають записи розмовного мовлення [див., напр.: Штейнфельдт 1963; Киссен 1964; Андрющенко 1968; Алексеев 1971; Saukkonen 1979; Дарчук 2005].

Під час підбору джерел до частотного словника, як правило, виявляється, що вирішити питання про вибірку з різних мовленнєвих жанрів і їхню питому вагу у словнику не можна, спираючись на хоч яку продуману теорію чи строгі аргументи, у будь-якому випадку неминучі довільні оцінки. Як результат – остаточне рішення за певним авторитетом у галузі лексикології [Засорина$_1$ 1966] або за оцінкою реальних можливостей конкретного автора чи колективу авторів [див.: Шубик 1980; Сутягина 1986; Алексеев 2001]. Іноді для забезпечення випадковості та однорідності вибірки використовують низку статистичних операцій [див.: Перебийніс 1985].

Дослідні можливості сучасних електронних текстових корпусів на сьогодні

загалом мають такі основні напрями [див.: Демська-Кульчицька$_1$ 2005]: 1) власне лінгвістичні синхронні та діахронні дослідження; 2) статистичні дослідження; 3) методика мови.

Під час укладання загальномовного корпусу також стоїть завдання зробити коректну вибірку, що базуватиметься на загальних і / або індивідуальних критеріях відбору емпіричного матеріалу [див.: Фрэнсис1983; Рабулец 2004; Шаров 2004; Meyer 2004; Демська-Кульчицька$_2$ 2005; Копотев 2006 та ін.]. Традиційно йдеться про розмір уривка, про кількість текстів із такою, а не іншою стилістично-жанровою, тематичною, структурною тощо характеристикою в корпусі та кількість слів у кожному текстовому фрагменті.

Джерелами укладання корпусів також правлять тексти різних стилів мовлення: художнього, наукового, офіційно-ділового, розмовного та публіцистичного [див., напр.: Демська-Кульчицька 2004; Касевич 2004; Meyer 2004]. Корпус містить по можливості усі типи писемних і усних текстів, репрезентованих у даній мові, усі ці тексти входять до корпусу по можливості пропорційно до їхньої частки у мові, а доброї репрезентативності досягають лише при значному обсязі корпусу (десятки і сотні мільйонів слововживань) [див.: Герд 2004; Захаров 2005; Корпусна... 2005; Копотев 2006].

На сьогодні домінує тенденція до складання корпусів з повних текстів. На основі повнотекстового корпусу досить легко побудувати традиційний корпус зразків [див.: Демська-Кульчицька$_1$ 2005, *61*; Корпусна... 2005, *32*]. Адже корпус текстів – це не просто колекція відібраних за певною методикою і презентованих у електронному вигляді текстів певних сфер вживання мови, а така колекція, яка категоризована як з боку інтегральних характеристик кожного тексту (наприклад, жанрових), так і з боку специфічних характеристик різних одиниць організації цього тексту (лексеми, словоформи, морфеми і т.ін.) [див.: Поликарпов 2007]. Така характеризація дозволяє провадити розгорнутий аналіз різного типу залежностей у текстах певної галузі.

Отже, способи відбору джерел для частотних словників і для корпусів природних мов суттєво між собою нічим не відрізняються. Впадає в око значно більший обсяг даних, презентованих у корпусі, порівняно з можливостями навіть великого частотного словника. Один із найбільших частотних словників, за редакцією Л.М. Засоріної, укладений на матеріалі 1 млн. слововживань [Частотный... 1977], тоді як найменший корпус мусить перевищувати 100 тис. слововживань.

Процедура укладання частотних словників передбачає послідовність таких етапів [див.: Засорина$_1$ 1966; Перебийніс 1985; Нелюбин 1991; Алексеев 2001]:

1) Вибрати одиницю підрахунку. Оскільки частотний словник укладають щоразу з певною метою, то з використанням корпусу цей етап не зміниться.

2) Укомплектувати вибірковий корпус текстів. Оскільки корпус укладають на різноманітній, добре продуманій (вимога репрезентативності – одна з основних ознак корпусу [див.: Демська-Кульчицька$_1$ 2005]) жанрово-тематичній базі, то вибірковий корпус текстів формують з уже створеного. Недоліком є певне нав'язування точки зору укладачів корпусу, проте незаперечна перевага у можливості обстежити усю генеральну сукупність, на яку фактично перетворюється заданий матеріал у корпусі.

3) Визначити експериментальний масив, мінімальний відрізок для статистичних підрахунків. З використанням корпусу така потреба відпадає. Якщо занадто не звужувати жанрово-тематичну структуру вибірки, то обсяг матеріалу в будь-якому разі перевищуватиме навіть найсміливіші побажання авторів частотних словників.

4) Вибрати спосіб підрахунку. На корпусному матеріалі спосіб підрахунку автоматичний.

5) Вирахувати частоту появи однакових одиниць підрахунку. З використанням корпусу це також виконують автоматично.

6) Врахувати один чи більше з трьох можливих видів омонімії (омографії): лексичну, лексико-граматичну чи граматичну. Відкритим для багатьох мов на сьогодні залишається питання визначення якісних ознак одиниць підрахунку майбутнього частотного словника. Це залежить передусім від якості розмітки корпусу, на основі якого формують частотний словник. Наприклад, графічній формі knd у перській

мові може відповідати третя особа однини теперішнього часу дієслова konad *робити*, третя особа однини теперішнього часу kanad і третя особа однини минулого часу kand дієслова *копати*. Зупинімося на цьому докладніше нижче, оскільки корпус текстів перської мови на сьогодні розмічений неповністю – усього декілька художніх творів.

7) Визначити статистичні характеристики одиниць частотного словника (коефіцієнт кореляції, відносну чи середню частоти та ін.). Це дозволяє зробити доступне нині усім програмне забезпечення Microsoft Excel. Перевірку надійності частотного словника здійснюють за допомогою поєднання функціональних можливостей зазначеного конкретного корпусу та загальнодоступного програмного забезпечення.

Розмітка сучасного корпусу полягає у приписуванні текстам і їхнім компонентам спеціальних поміток: зовнішніх (екстралінгвістичних) і власне лінгвістичних, які описують лексичні і граматичні характеристики елементів тексту. Набір цих метаданих багато у чому визначає можливості, які корпус надає дослідникам [див.: Герд 2004]. Як лексична (лематизація), так і морфологічна розмітка спрощує пошук і лінгвістичну роботу з корпусом. Особливої ваги вона набуває для мов з високим ступенем флективності [див.: Копришивова 2004]. Оскільки перська мова аналітичного типу [див.: Бекбаева 1982], то незавершена розмітка корпусу вплине лише на розмежування омонімів (точніше, омографів).

Проблема розмежування омографів у частотному словнику (та й корпус також почали розмічувати з метою її розв'язання) постала уже давно. Під час ручного розписування текстів цю проблему вирішують відносно просто – укладач завжди має перед собою контекст. Однак на практиці частотних семантичних словників, які б реєстрували лексичні значення вхідних одиниць, існує небагато. Вони складають окрему галузь статистичної лексикографії. Лише окремі частотні словники вказують лексико-граматичні значення своїх вхідних одиниць. У них або кожне слово, або тільки ті слова, які збігаються за написанням, одержують індекс частини мови. Майже всі іноземні частотні словники не повідомляють ніякої інформації стосовно належності слів до певних частин мови чи граматичних категорій, тобто не враховують ні лексико-граматичних, ні граматичних значень [див.: Засорина$_1$ 1966; Алексеев 2001; Arabic... 2007].

Для розпізнавання омографів у нерозміченому корпусі треба або розробляти складні алгоритми, або підраховувати вручну на основі конкордансів. Основна маса загальновживаних слів буде розпізнана надійно, а решту слів, які складають меншу частку словника будь-якого тексту, можна контролювати шляхом перегляду в конкордансі і прийняття рішень у кожному випадку окремо [див.: Фрэнсис 1983; Использование... 1990]. Також ефективним для зняття омографії буде використання макро- і мікроконтекстів [див. дет.: Зубов 1971]. Для перського корпусу така процедура видається цілком реальною – опція укладання конкордансів доступна для усіх текстів.

Доступний в Інтернеті [http://pldb.ihcs.ac.ir/] корпус текстів перської мови укладений на базі сучасної перської мови – мови останніх 75 років [див.: Āssi$_1$ 2008]. Його можна визначити як повнотекстовий, моніторинговий (динамічний), синхронний, загальномовний, літературний, частково лінгвістично анотований (розмічений), змішаний за типом реалізації мовної системи, обмежено доступний. Потенційно (відповідно до декларації укладачів [див.: Assi 2008]) дослідницький, але на сьогодні фактично ілюстративний. Розмітка корпусу передбачає фонетичну, морфологічну, граматичну та семантичну інформацію про кожне слово. З трьох запланованих етапів розробки корпусу [див.: Āssi$_2$ 2008] на сьогодні втілюється другий. Цей корпус містить як писемну мову, так і розмовну, мову різних стилів і різні субмови. За жанровим розподілом близько половини текстів становить художня література. Решта текстів належать до публіцистичного та наукового стилів, також є зразки розмовного стилю. Дослідник має змогу укласти частотний список на базі зразків як розмовної (виклад новин), так і писемної мови (прози і поезії), різних стилів (художнього, розмовного, публіцистичного) і різних субмов (медицини, спорту, мистецтва і т.ін.). Можна вибрати усі доступні опції, а можна

укладати частотний список за субкорпусами одного типу. Частотний словник може бути за спадом або за зростанням частот, у прямому або зворотному алфавітному порядку. Також можна обмежити кількість слів (це актуально передусім для лінгводидактичних потреб).

Укладачі корпусу перської мови використали по можливості всі типи письмових текстів, презентованих у мові, усні тексти, правда дещо обмежено, також представлені. Корпус у цілому налічує 1500 повних текстів. Під час укладання частотних списків подають усю необхідну статистичну інформацію: кількість слів у тексті, кількість словоформ, відносна й абсолютна частота слова, його ранг. Інші статистичні дані можна вирахувати на підставі поданих.

Складні дієслова тут подають як два (чи більше) окремі слова. Власні назви включають на загальних підставах як окремі слова. За бажанням їх майже завжди можна виокремити і вилучити.

За допомогою спеціалізованого програмного інструментарію за матеріалами будь-якого корпусу можна укладати словники, індекси і конкорданси [див.: Белоногов 2004; Корпусна... 2005; Поликарпов 2007]. „У принципі, остаточний результат не має залежати від того, чи використовують ручне розписування тексту, чи машинний аналіз. Але в людини неминучі помилки під час монотонної роботи. Природними є помилки в арифметичних підрахунках; щоб їх виправляти, доводиться у найвідповідальніших випадках виконувати контрольні підрахунки" [Алексеев 2001, *60–61*]. Правда, поки що повністю автоматизувати процес укладання частотного словника можна лише для добре опрацьованих мов, таких як, наприклад, російська [див.: Поликарпов 2007] чи англійська [див.: Meyer 2004]. Програмне забезпечення і стан вивчення перської мови тим часом цього зробити не дозволяють, проте ми констатуємо існування принципової можливості укладання частотного словника на матеріалі корпусу перських текстів. Корпус є репрезентативним для поставленої мети – укладання частотного загальномовного чи галузевого словника. У перспективі – дослідження із лексичної семантики, семантичної сполучуваності та укладання на базі корпусу перської мови частотного семантичного словника.

ЛІТЕРАТУРА

Алексеев П.М. **Частотные словари.** Санкт-Петербург, *2001*.

Алексеев П.М. Частотные словари английского языка и их практическое применение // **Статистика речи и автоматический анализ текста.** Ленинград, *1971*.

Андрющенко В.М. Новые работы в области статистической лексикографии // **Вопросы языкознания,** *1968*, №5.

Арнольд И.В. **Основы научных исследований в лингвистике.** Москва, *1991*.

Бекбаева К.А. Персидский язык // **Квантитативная типология языков Азии и Африки.** Ленинград, *1982*.

Белоногов Г.Г. **Компьютерная лингвистика и перспективные информационные технологии.** Москва, *2004*.

Бук С. Частотний словник офіційно-ділового стилю: принципи укладання та статистичні характеристики лексики // **Лінгвістичні студії,** *2006*, №14.

Герд А.С., Захаров В.П. Нерешенные вопросы Национального корпуса русского языка // **Международная конференция "Корпусная лингвистика – 2004". Тезисы докладов.** Санкт-Петербург, *2004*.

Дарчук Н. Параметризована база даних сучасної української мови на основі частотних словників // **Проблеми квантитативної лінгвістики**, Чернівці, *2005*.

Демська-Кульчицька О.М. Національний корпус української мови // **Вісник Київського національного лінгвістичного університету. Серія "Філологія"**, *2004*, т. 7, №1.

Демська-Кульчицька О. **Основи Національного корпусу української мови.** Київ, *2005*.

Демська-Кульчицька О. Репрезентативність як ознака текстового корпусу // **Українська мова,** *2005*, №3.[2]

Засорина Л.Н. **Автоматизация и статистика в лексикографии.** Ленинград, *1966*.[1]

Засорина Л.Н. Частотные словари и вопросы лексикостатистики // **Межвузовская конференция по вопросам частотных словарей и автоматизации лингвистических работ. Тезисы докладов и сообщений.** Ленинград, *1966*.[2]

Захаров В.П. **Корпусная лингвистика.** Санкт-Петербург, *2005*.

Зубов А.В. Переработка текста естественного языка в системе «человек – машина» // **Статистика речи и автоматический анализ текста.** Ленинград, *1971*.

Использование ЭВМ в лингвистических исследованиях / *Т.А. Грязнухина, Н.П. Дарчук, Н.Ф. Клименко и др.; Отв. ред. В.И. Перебейнос.* Киев, *1990*.

Касевич В.Б. О работе по созданию Национального корпуса русского языка в Лаборатории моделирования речевой деятельности СПбГУ // **Международная конференция "Корпусная лингвистика – 2004". Тезисы докладов.** Санкт-Петербург, *2004*.

Киссен И.А. Опыт статистического исследования частотности лексики передовых статей газеты «Кизил Узбекистон» // **Научные труды Ташкентского государственного университета им. В.И. Ленина. Филологические науки.** Ташкент, *1964*.

Копотев М.В., Янда Л. Национальный корп ус русского языка (www.ruscorpora.ru) // **Вопросы языкознания**, *2006*, №5.

Копришивова М. К некоторым вопросам, связанным с лемматизацией корпуса чешских текстов // **Международная конференция "Корпусная лингвистика – 2004". Тезисы докладов.** Санкт-Петербург, *2004*.

Корпусна лінгвістика / *В.А. Широков, О.В. Бугаков, Т.О. Грязнухіна та ін.* Київ, *2005*.

Малаховский Л.В. О принципах построения частотных словарей английского языка // **Межвузовская конференция по вопросам частотных словарей и автоматизации лингвистических работ. Тезисы докладов и сообщений.** Ленинград, *1966*.

Мартинюк С. Вживання частотних словників у вивченні східних мов // **Мовні і концептуальні картини світу,** *2003*, №9.

Муравицька М.П. Статистичні лінгвістичні дослідження та їх розвиток в українському мовознавстві // **Мовознавство,** *1967,* №5.

Нелюбин Л.Л. **Компьютерная лингвистика и машинный перевод**. Москва, *1991*.

Перебийніс В.І. **Статистичні методи для лінгвістів**. Вінниця, *2002*.

Перебийніс В.С., Муравицька М.П., Дарчук Н.П. **Частотні словники та їх використання**. Київ, *1985*.

Поликарпов А.А. **Компьютерный корпус текстов русских газет конца XX-ого века** // http://www.philol.msu.ru/~lex/corpus/corp_descr.html (скопійовано 01.04.2008).

Прикладна лінгвістика та лінгвістичні технології: MegaLing-2006: Зб. наук. праць / За ред. *В.А. Широкова*. Київ, *2007*.

Рабулец А.Г., Костышин А.М., Сидорчук Н.М., Широков В.А. Системотехнические и лингвистические принципы проектирования украинского лингвистического корпуса // **Международная конференция "Корпусная лингвистика – 2004". Тезисы докладов.** Санкт-Петербург, *2004*.

Сутягина Л.М. Формирование выборочного корпуса текстов при составлении частотного словаря (качественная сторона формирования выборки) // **Квантитативные методы отбора учебного материала по иностранному языку для неязыкового вуза.** Свердловск, *1986*.

Тулдава Ю. **Проблемы и методы квантитативно-системного исследования лексики.** Таллин, *1987*.

Троянская Е.С. Отбор языкового материала для статистической обработки // **Филологические науки**, *1967*, №3.

Фрэнсис У.Н. Проблема формирования и машинного представления большого корпуса текстов // **Новое в зарубежной лингвистике,** *1983*, №14.

Частотный словарь русского языка. Около 40 000 слов / *Л.Н. Засорина*. Москва, *1977*.

Шаров С.А., Савчук С.О. Типология текстов для представительного корпуса // **Международная конференция "Корпусная лингвистика – 2004". Тезисы докладов.** Санкт-Петербург, *2004*.

Штейнфельдт Э.А. **Частотный словарь современного русского литературного языка.** 2500 наиболее употребительных слов. Таллин, *1963*.

Шубик С.А. Статистические методы в лингвистике // **Статистика речи и автоматический анализ текста.** Ленинград, *1980*.

Arabic Word Frequency Counts // http://www.qamus.org/wordlist.htm (скопійовано 01.04.2008).

Āssi M. **Pāygāhdādehā-ye zabān-e fārsi dar internet** // http://pldb.ihcs.ac.ir/Files/PLDB-REP.pdf (скопійовано 01.04.2008).[1]

Āssi M. Tarh-e ijād-e pāygāhdādehā-ye zabān-e fārsi bā komak-e kāmpyuter. // **Ettelā'rasāni-ye dowre-ye 11 shomāre-ye 1 – zemestān 1373** // http://www.irandoc.ac.ir/ETELA-ART/11/11-1-2.htm (скопійовано 01.04.2008).[2]

Assi S.M., Abdolhosseini M.H. **Grammatical Tagging of a Persian Corpus** // http://pldb.ihcs.ac.ir/Files/Assi.pdf (скопійовано 01.04.2008).

Bybee J., Hopper P. (Eds.) **Frequency and the Emergence of Linguistic Structure.** Amsterdam, *2001*.

http://pldb.ihcs.ac.ir/ (скопійовано 01.04.2008).

Meyer Ch. **English Corpus Linguistics. An introduction.** Cambridge, *2004*.

Saukkonen P., Haipus M., Niemikorpi A., Sulkala H. **Suomen kielen taajuussanasto. A Frequency Dictionary of Finnish.** Porvoo, Helsinki, Juva, *1979*.