Старко В.Ф., к. філол. н., доц.,
Східноєвропейський національний університет
імені Лесі Українки

ФОРМУВАННЯ БРАУНСЬКОГО КОРПУСУ УКРАЇНСЬКОЇ МОВИ

У статті викладено принципи побудови й завдання Браунського корпусу української мови (БрУК) та розглянуто проблеми його текстового наповнення. Пояснено застосування програми LanguageTool для оцінки якості текстів.

Ключові слова: *корпус, побудова корпусу, БрУК, українська мова, корпусна лінгвістика, LanguageTool, перевірка тексту.*

Браунський корпус (Standard Corpus of Present-Day Edited American English, або скорочено Brown Corpus), що його створили В. Нельсон Френсис та Генрі Кучера в Браунському університеті в 1960-х роках, досі слугує взірцем для створення подібних корпусів-мільйонників, зокрема й для інших мов. Переваги створення "родини" таких корпусів полягають у збалансованості, можливості порівняльних досліджень, їх використанні в курсах із корпусної лінгвістики тощо [Старко 2012], [Cheilytko 2013]. З цих міркувань Браунський корпус було взято за модель для корпусу української мови. Наші корпусні дослідження на українському матеріалі [Starko 2013], [Старко 2014] засвідчили плідність застосування корпусних методів у вивченні української мови й водночас засигналізували потребу у використанні якомога збалансованішого й репрезентативнішого корпусу. Важливість репрезентативності, збалансованості, докладної параметризації корпусу ми обґрунтували в окремій праці [Старко 2013], а в проекті БрУК їх було взято за методологічні засади. В роботі над БрУК ми активно користуємося працями Ніни Федорівни Клименко, Євгенії Анатоліївни Карпіловської та інших працівників відділу структурно-математичної лінгвістики Інституту української мови НАНУ [Активні 2013], [Граматичний 2011], [Клименко 2008]. Серед них особливе місце посідає "Граматичний словник української літературної мови" – настільна книга кожного, хто береться за автоматизований морфологічний аналіз української мови.

На початку розроблення БрУКу ми сформулювали такі вимоги до текстів:

1. Оригінальні (тобто неперекладні) твори.

2. Твори, створені й опубліковані за відносно короткий проміжок часу.

3. Тексти мають бути зредаговані.

4. Тексти слід підібрати згідно з класифікацією за категоризаційною матрицею первісного Браунського корпусу, що має 15 категорій, поділених на різну кількість підкатегорій [Francis 1979].

5. Корпус має складатися з 500 фрагментів довжиною 2000 слів.

6. Фрагмент корпусу повинен бути взятий із одного тексту, окрім випадків, коли тексти дуже короткі, як-от новини.

Первісно нам ішлося про те, щоб максимально відтворити структуру американського корпусу, але водночас врахувати українську специфіку. Кількість та розмір категорій текстів, що складають Brown Corpus, його творці визначили шляхом усереднення даних опитування учасників конференції [Francis 1979], а докладніше розбиття на підкатегорії відбивало пропорційні співвідношення друкованих видань у США за 1961 рік. Як виявилося пізніше, такий мовний зріз важко відтворити на матеріалі британського варіанта англійської мови й перенести його на інші часові відтинки. Суперечність між дотриманням пропорцій Brown Corpus та прагненням репрезентативно представити стан англійської мови в іншій країні чи в інший час неодмінно змушувала розробників до компромісів і відступів. Стосовно інших мов ця суперечність ще більше загострюється. Наприклад, у первісному корпусі фігурує жанр "вестерн", якого годі знайти в слов'янських мовах.

У процесі роботи над проектом БрУК еволюціонувало наше бачення завдань корпусу та його структури. За мету було покладено створення "взірцевого корпусу" – взірцевого в сенсі якості текстів. До такого рішення нас спонукала мовно-соціальна ситуація в Україні. З одного боку, впав загальний мовний рівень творів, чимало текстів, особливо, онлайнових, публікують без належного редагування та з численними помилками. І в електронному, і в друкованому форматах

поширені неявні переклади з російської мови (інколи з помітними ознаками машинного перекладу) – цим грішать навіть великі видавництва й потужні ЗМІ. З другого боку, чимало мовців прагнуть підвищувати свою культуру мовлення, шукають взірців добірної української мови, прагнуть орієнтуватися на найкращі в мовному плані твори. Застосування суто статистичного й дескриптивного підходу без жодного контролю за якістю й походженням тестів може призвести до захаращення корпусу третьосортними текстами. З цих міркувань ми поклали добирати тексти щонайвищої якості.

Визнаючи потребу адекватно відбити в корпусі особливості функціювання саме української, а не англійської, мови та передбачаючи можливість порівняльних досліджень із іншими, віддаленими часовими зрізами, ми вирішили укрупнити категорії Браунського корпусу, але не виходити за межі загальної його структури. Загалом тексти в корпусі поділяються на два види – інформаційні (призначені поінформувати читача) та художні (описують вигаданих персонажів і події). У БрУК інформаційний вид охоплює категорії A-H, художній – лише I. Таким чином, із 15 категорій Brown Corpus було отримано 9. Обсяги кожної з них було визначено шляхом усереднення індивідуальних оцінок, що їх дали учасники проекту:

A. Преса – 25%.

B. Релігійна література – 3%.

C. Професійно-популярна література ("сад і город", ремонт, ремесла, хобі, робочі професії тощо) – 7%.

D. Інформаційні тексти (що не потрапляють в інші категорії, зокрема біографії, мемуари, есеї тощо) – 7%.

E. Адміністративні документи (закони, урядові акти, звіти організацій тощо) – 3%.

F. Науково-популярна література – 5% .

G. Наукова література – 10%.

H. Навчальна література – 15%.

I. Художні тексти – 25%.

В межах кожної категорії забезпечується тематичне, жанрове, географічне й авторське розмаїття, а порівнювати корпуси можна буде за категоріями.

Щоб збільшити гнучкість у підбиранні текстів, ми відмовилися від ідеї формування фрагментів корпусу у вигляді суцільних витягів строго довжиною 2000 слововживань. Натомість уривки текстів можуть бути й коротші, аби лише в сумі набирався згаданий обсяг із одного джерела.

Наразі ми формуємо БрУК-2013, тобто основний масив текстів датовано 2013 роком, але для невеликої кількості текстів допускаємо відхил на два (у виняткових випадках – три) роки. Основним критерієм служить дата першої публікації твору, а другорядним – дата написання. Наприклад, не розглядаємо твір, вперше виданий 1990 року й перевиданий 2013 року, і текст, створений 1975 року, але вперше опублікований у 2012 році. Натомість якщо автор писав свій твір упродовж кількох років і видав його 2013 року, він задовольняє часовий критерій.

Услід за Brown Corpus БрУК не містить віршів (за винятком коротких уривків, цитованих у прозових творах), драм (на тій підставі, що це насправді не письмовий дискурс, а художнє відтворення розмовного дискурсу) й уривків із прозових творів, що на понад 50% складаються з діалогів. Мова української діаспори становить окремий важливий об'єкт дослідження й до БрУКу не входить.

Повищі вимоги достатньо формалізовані й надаються до адекватного контролю. Виняток становить хіба що вимога оригінальності (неперекладності) твору. Річ у тім, що деякі українські видавництва видають ту саму книжкову продукцію (наприклад, збірки рецептів, навчальну літературу) одночасно українською й російською мовами без зазначення вихідної мови твору. Низка українських ЗМІ мають паралельні українську й російську версії, через що буває важко встановити первісну мову тієї чи тієї публікації. У випадках, коли є обґрунтований сумнів щодо первісної мови твору, він не потрапляє до корпусу.

Значно важче формалізувати вимогу щодо високої якості текстів. Укладачі намагаються добирати тексти яскраві, нешаблонні, розкуті, багаті лексикою тощо. Виправляємо лише абсолютно очевидні, суто друкарські помилки (наприклад, **мадуть* замість *мабуть*). Інші помилки документуємо в спеціальній зоні метаопису фрагмента корпусу. Контролювати

якість матеріалу допомагає спеціальна програма перевірки українських текстів [LanguageTool 2014], що її розробив один із учасників проекту. Цей програмний засіб провадить граматичну й стильову перевірку текстів і має ширший функціонал проти стандартних програм перевірки орфографії. Ним можна скористатися у формі додатку до офісних програм, окремої програми й веб-служби (за адресою r2u.org.ua/check). Програма оперує низкою граматичних та стильових правил і повідомляє про орфографічні й граматичні помилки, порушення правил сполучуваності, варваризми, а також пропонує правильний чи доречніший спосіб висловлювання. В окремому файлі зібрано правила перевірки пунктуації, узгодження, милозвучності тощо. Наприклад, ці правила виловлюють такі хибні вживання, як **згідно статуту*, *три вулика*, *20 байт*, невиправдані кальки (**в деякій мірі*, *приймати участь*, *в міру необхідності*, *піти за хлібом*, *вести себе* тощо), порушення милозвучності (**в вівторок*), невиділення комами вставних слів а також менш очевидні огріхи на кшталт поєднання пасивного предиката з орудним відмінком діяча (**Справу порушено прокурором*). Кілька сот помилок охоплено простими правилами заміни за шаблоном "неправильний вираз=пропоновані варіанти", наприклад:

багатообіцяючий=багатонадійний|багатообіцяльний
всерйоз=серйозно|навсправжки
енергозбереження=енергоощадження
недомогання=нездужання|недугування|легка недуга|легка неміч
обнуління=занулення|занулювання.

Коли спрацьовує правило, програма пропонує заміну й інколи відсилає користувача до джерела зі стилістики мови, наприклад:

More info: http://yak-my-hovorymo.wikidot.com/porivnyannya-u-porivnyanni-porivnyano-yak-porivnyaty-proty
Message: Краще: 'коштом', 'шляхом', 'користуючись з', 'ким/чим', 'внаслідок'
... Механічне розширення складу партії за рахунок сателітів

Засіб LanguageTool та його правила перевірки перебувають на етапі дальшого розширення й вдосконалення, проте вже зараз

наявний функціонал дає змогу швидко оцінити якість тексту в електронному форматі.

Отже, будова Браунського корпусу української мови в загальних рисах спирається на модель первісного Браунського корпусу. Критерії, висунуті до текстів, відбивають прагнення авторів створити взірцевий у мовному плані корпус сучасної української мови з дотриманням принципів репрезентативності, збалансованості й порівняльності. У роботі над корпусом укладачам стали в пригоді праці Н.Ф. Клименко та її колег, а також програма-коректор LanguageTool.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Активні ресурси сучасної української номінації: Ідеографічний словник нової лексики / [Карпіловська Є.А., Кислюк Л.П., Клименко Н.Ф. та ін.]; відп. ред. Є.А. Карпіловська. – К.: ТОВ "КММ", 2013. – 416 с. 2. Граматичний словник української літературної мови. Словозміна: Близько 140 000 слів / [Критська В.І., Недозим Т.І., Орлова Л.В. та ін.]; відп. ред. Н.Ф. Клименко. – К.: Вид. дім Дмитра Бураго, 2011. – 760 с. 3. Клименко Н.Ф. Динамічні процеси в сучасному українському лексиконі: Монографія / Клименко Н.Ф., Карпіловська Є.А., Кислюк Л.П. – К.: Вид. дім Дмитра Бураго, 2008. – 336 с. 4. Старко В. Концепція створення Браунського корпусу української мови / Старко В., Чейлитко Н. //"Комп'ютерна лінгвістика: сучасне та майбутнє". Матеріали Міжнародної науково-практичної конференції – К.: КНЛУ, 2012.– С. 45-46. 5. Старко В. Корпусні дані в дослідженні українських колоративів / Старко В. // Українська мова. – 2014. – Вип. 1. – С. 51-60. 6. Старко В. Параметризація корпусу як спосіб підвищити його репрезентативність та збалансованість / Старко В., Чейлитко Н. // Українське мовознавство. – Вип. 43. – Київ, 2013. – С. 87-94. 7. Cheilytko N. The Ukrainian Brown Corpus and Dependency Tree Modeling / Cheilytko N., Starko V., Galkin A. // Досвід розробки та застосування приладо-технологічних САПР в мікроелектроніці: Матеріали XII Міжнародної науково-технічної конференції CADSM 2013. – Львів: Вид-во Нац. ун-ту "Львівська політехніка", 2013. – С. 58-60. 8. Francis W.N. Brown Corpus Manual / Francis W.N., Kucera H. – Providence, Rhode Island: Brown University, 1979. – http://icame.uib.no/brown/bcm.html – режим доступу: 24.09.2014 p. 9. LanguageTool – r2u.org.ua/languagetool/about – режим доступу: 24.09.2014p. 10. Starko V. Ukrainian Colour Concepts for Blue / Vasyl Starko // Slovo. Journal of Slavic Languages and Literatures, 2013. – No. 54. – Pp. 150-163.

Старко В.Ф., к. филол. н., доц.,
Восточноевропейский национальный университет
имени Леси Украинки

ФОРМИРОВАНИЕ БРАУНОВСКОГО КОРПУСА УКРАИНСКОГО ЯЗЫКА

В статье излагаются принципы построения и задания Браунского корпуса украинского языка (БрУК) и рассматриваются проблемы его текстового наполнения. Объясняется использование программы LanguageTool для оценки качества текстов.

Ключевые слова: *корпус, построение корпуса, БрУК, украинский язык, корпусная лингвистика, LanguageTool, проверка текста.*

Starko V., PhD., Assistant Professor,
Lesya Ukrainka Eastern European National University

DEVELOPMENT OF BROWN CORPUS OF THE UKRAINIAN LANGUAGE

The article sets forth the design principles and the goal of the Brown Corpus of the Ukrainian language (BCU) and discusses problems of text selection for the corpus. The use of the Language Tool software in evaluating the quality of texts is specified.

Keywords: *corpus, corpus design, BCU, the Ukrainian language, corpus linguistics, Language Tool, spellchecker.*

УДК 811(112.2+113.6)

Стасюк О.С., к.філол.н., асист.,
Інститут філології КНУ імені Тараса Шевченка

ІНТЕРТЕКСТУАЛЬНІСТЬ ЯК ЗАСІБ ДИСКРЕДИТАЦІЇ ПОЛІТИЧНИХ ОПОНЕНТІВ
(на матеріалі парламентської комунікації ФРН і Швеції)

У статті визначено прагматичні особливості використання інтертекстуальних включень у парламентській комунікації ФРН і Швеції з метою дискредитації політичних опонентів. Виокремлено типи інтертекстуальних включень, які використовуються для дискредитації політичних опонентів у парламентській комунікації,