**Критерії підбору текстів для БрУК**

Браунський корпус української мови (БрУК) задуманий так, що в певних рисах він повторюватиме структуру оригінального Браунського корпусу й взагалі спиратиметься на «ідеологію» цього корпусу.

**Вимоги до текстів:**

* оригінальні (тобто неперекладні)
* прозові
* зредаговані (учнівський твір чи приватний лист не годиться)
* написані (й опубліковані) переважно в 2013 році. Якщо таких нема, то беремо за плюс/мінус 2 роки (тобто 2011-2014рр.). Як виняток - 2010 рік. Твір, написаний, скажімо, 1995 чи 2008 року, а виданий 2013 року, не підходить.
* кожен текст має бути паспортизований (перелік метаданих нижче)
* кожен текст вміщуємо в окремому текстовому файлі
* до корпусу беремо зазвичай **один уривок** довжиною 2000 слововживань + ще кілька слововживань до кінця поточного речення. Якщо текст (наприклад, коротке інформаційне повідомлення) містить менше за 2000 слововживань, подаємо його цілком в окремому файлі

**Вимоги до якості текстів:**

* без найменших ознак машинного чи людського перекладу (чимало текстів на pravda.com.ua та інших новинних сайтів цієї вимоги не задовольняють)
* грамотність - відсутність орфографічних, пунктуаційних, граматичних помилок тощо
* стиль - відсутність русизмів, активних дієприкметників, невдалих синтаксичних конструкцій (типу *План виконується нами*)
* намагаємося підбирати тексти нешаблонні, жваві, розкуті, багаті лексикою, з влучними формулюваннями (див. приклади)

Перевірити електронний текст на наявність орфографічних і стилістичних помилок можна тут:

r2u.org.ua/check - утилітою LanguageTool. Окрім однозначних помилок програма сигналізує про невдалі варіанти й рекомендує кращі. Пропоновані поправки слід сприймати критично.

**Джерела текстів:**

* електронні тексти з відкритих інтернет-джерел - газетні й журнальні публікації, уривки художньої та іншої літератури, наукові статті
* скани, викладені в інтернеті
* самі скануємо паперові видання (кілька сторінок), розпізнаємо й вичитуємо потрібної довжини фрагмент

**Категорії текстів у БрУК**

А. Преса (репортажі, огляди, редакційні статті, листи до редакції; національні й регіональні видання; тематично - політика, спорт, суспільство, економіка й фінанси, короткі новини, культура - театр, література, музика, танці)

B. Релігійна література (книжки, періодика, брошури).

С. Професійно-популярна література (книжки й періодика; домоводство, ремесла, «сад і город», хобі, ремонт і будівництво, конструювання, хобі, музика й танці, домашні тварини, спорт, їжа й вино, подорожі, фермерство, робочі професії тощо).

D. «Естетичні інформативні» тексти (інформативні тексти, що не потрапляють в інші категорії, зокрема, біографії, мемуари, есеї, особисті листи, художня й мистецтвознавча критика)

E. Адміністративні документи (закони, урядові акти, звіти організацій/фондів/компаній, офіційні листи)

F. Науково-популярна література.

G. Наукова література (книжки й періодика; природничі й гуманітарні науки, техніка й інженерна справа).

H. Навчальна література (підручники, посібники тощо, гум. і прир. науки та інше)

I. Художні тексти (романи, повісті, оповідання, новели, за тематикою – загальна, детективи, фантастика, пригодницька, любовна, гумористична тощо)

Загалом тексти поділяються на два види – інформативні (призначені поінформувати читача) і художні (описують вигаданих персонажів і події). Інформативний вид охоплює категорії А-H, художній – лише I.

**Метадані**

Наразі метадані до текстів подаємо першим рядком у файлі з текстом у такому форматі:

A002 Макаров Юрій. Dura lex Український тиждень Ч.32 2013 tyzhden.ua/Columns/50/86563 622

де:

А002 – ідентифікатор тексту, що вказує на категорію текстів (А) і порядковий номер тексту (002).

Макаров Юрій - прізвище й ім’я автора тексту

Dura lex - назва тексту

Український тиждень - назва друкованого видання

Ч.32 - число періодичного видання

2013 - рік публікації

tyzhden.ua/Columns/50/86563 - адреса, звідки взято текст; для паперових книжок - місто, видавництво, сторінки

622 - кількість словоформ у тексті

Потрібно ще буде ввести поле «Помічені помилки», в якому документувати помилки й подавати виправлені варіанти. Досвід роботи з текстами показує, що «ідеальних» текстів не так вже й багато. Навіть найкращі тексти можуть мати певні хиби.

Раніше розробили були докладнішу паспортизацію, але для БрУКу вона надмірна. Серед найважливіших додаткових типів даних: тема й стиль. Варто передбачити можливість додавати нові типи даних до метаопису, наприклад, якщо БрУК увійде до більшого корпусу з докладнішою параметризацією.

**Відсоткові частки**

преса – 25%

релігійна – 3%

проф-популярна – 7%

естетична – 7%

адміністративні – 3%

науково-популярна – 5%

наукова – 10%

навчальна – 15%

худ. тексти – 25%