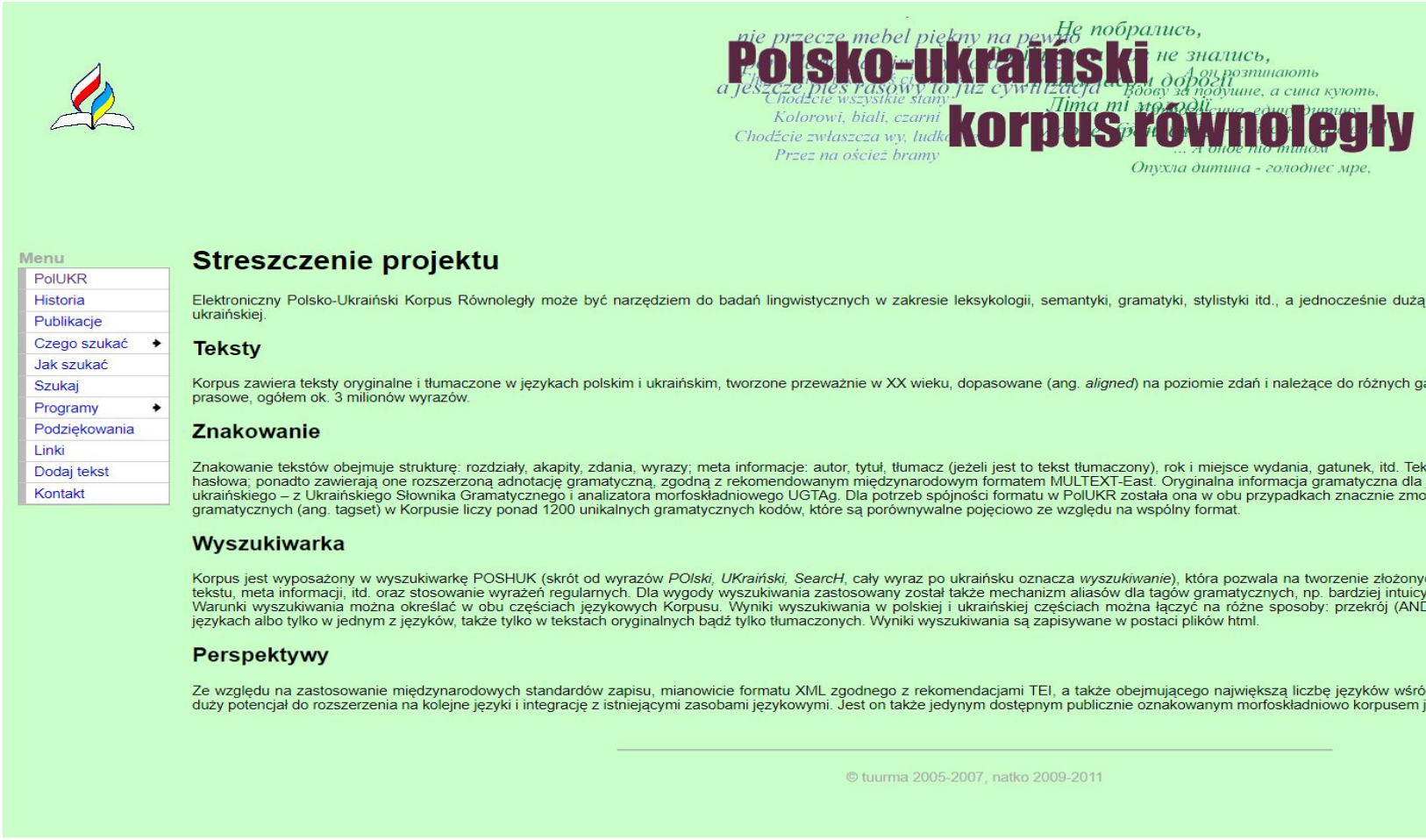
POLSKO-UKRAIŃSKI KORPUS RÓWNOLEGŁY



Корпус містить оригінальні та перекладені тексти польською та українською мовами, створені здебільшого у 20 столітті, вирівняні на рівні речень та належать до різних жанрів: художньої літератури, публіцистики, підручників, документів, прес-релізів, загалом близько 3 мільйонів слів.

ІСТОРІЯ ПРОЕКТУ

АВТОРИ ПРОЕКТУ

Наталія Коциба

Магдалина Турська

Ідея народилася на неофіційній основі як експеримент. Він був натхненний сесією Міжнародної гуманітарної школи, присвяченій корпусній лінгвістиці, організованій у Міжпредметному науково-дослідному інституті Варшавського університету в січні 2004 р. Ця сесія була призначена для молодих дослідників Центральної та Східної Європи. Там було показано можливості використання корпусу польської мови в лексикографічному та лінгвістичному дослідженнях загалом. На жаль, для української мови таких ресурсів на той час у відкритому доступі не було, не кажучи вже про двомовних. З іншого боку, відчувається відсутність великого сучасного польсько-українського словника, який частково міг би перейняти паралельний корпус цих мов.

У листопаді 2004 року почали збирати тексти. У квітні 2005 року з'явилася перша концепція корпусу, а вже у вересні - його пілотна версія. Він містив 50 невеликих текстів (25 пар), переважно публіцистичних, отриманих від перекладачів. Ці тексти були вирівняні на рівні абзацу і містили основні метаінформації: назва, автор, перекладач, мова оригіналу тощо.

З жовтня 2007 року проект отримав дворічну фінансову підтримку Міністерства інформаційних технологій та вищої освіти Республіки Польща, що змінило свій неофіційний статус та дозволило йому розвиватися далі.

Ось найважливіші зміни:

Корпус був значно розширений - наразі він містить понад 3 млн слів.

Підходить на рівні речень, а не абзаців, як раніше.

Тексти лематизовані та містять синтаксичну інформацію морфа, при цьому набори тегів для польської та української є стандартизованими.

Корпус оснащений пошуковою системою POSHUK, яка дозволяє поєднувати параметри різних рівнів маркування (структурний, морфа-синтаксичний, метейнформаційний, а також одночасне налаштування параметрів пошуку на обох мовах.

Корпус можна використовувати не тільки в Інтернеті, але і встановивши його на локальний комп'ютер (ця опція з’явиться незабаром)

Зараз ведеться робота над розрізненням синтаксичного маркування морфів для українських текстів. Плани на найближче майбутнє також - збагатити тексти смисловою інформацією.

ЯК ЗДІЙСНЮВАТИ ПОШУК

kropka .	zamienia **dowolny znak** zapytanie [word="b.g"] daje wyniki z wyrazami *big bag beg bog bug*
gwiazdka *	po kropce znajduje **dowolną liczbę dowolnych liter** .*iwać znajdzie wyrazy, któe kończą się na *-iwać* przy.* znajdzie wyrazy, któe zaczynają się na *przy-* bl.*ng znajdzie wyrazy, któe zaczynają się na *bl-* i kończą się na *-ng* .*sta.* znajdzie wyrazy, które mają w środku ciąg *sta.*
pytajnik ?	Znaki w nawiasach są opcjonalne. blond(e)? znajdzie *blond* oraz *blonde*
kreska pionowa |	*walka|bitwa|wojna* szuka dowolnego z podanych wyrazów
nawiasy kwadratowe [i]	Mogą zawierać znaki pojedyńcze alternatywne. «[rґ]рати» znajdzie dwa warianty pisowni *«ґрати» i «rґрати»*

У запитах можна шукати значення наступних атрибутів: лема, слово, тег. Кожен елемент запиту повинен бути укладений у квадратні дужки: [], запит може містити багато таких елементів. Значення атрибутів пошуку слід розміщувати в лапках: "", наприклад [lemma = "день"] або [word = "this"] або [tag = "Spg"].

Ви можете використовувати наступні оператори у своїх значеннях пошуку

Великі літери

Якщо ви введете доктора в лемі, ми отримаємо лише результати лікаря.

Але лікар як лема дає всі можливості, малі та малі літери.

Щоб отримати лише лікаря, ви повинні додати прапор / i після сегмента, наприклад [lemma = "doctor"] / i.

[word = "Варшава | Краків"]

[слово = "зелений | синій | жовтий"]

Можливий пошук певної частини мови або іншої визначеної морфологічної інформації.

Список частин мови, змінні: N (іменник), A (прикметник), V (дієслово), R (прислівник), P (займенник), M (числівник) та незмінний: S (прийменник), C (сполучник), I (знак оклику), Q (частинка) та дві технічні категорії: Y (абревіатура, абревіатура), X (невизнаний, залишковий).

Див. детальний опис для польської та української

Синтаксис запиту:

[tag = ""]

Примітка. Ви можете використовувати крапку в тегах синтаксису морфа. Наприклад, всі дієслова починаються з V.
На другому місці - інформація про тип дієслова, лексичну та допоміжну "бути".
Третє місце виділяє аспект: p недосконалий (прогресивний) та досконалий. Четверте місце займає інформація про форму дієслова: i - вказівний режим, c - умовний режим, m - імперативний режим, n - інфінітивний, o - безособова форма (на -o форма), г - дієприслівниковий дієприкметник (герундія).
Наприклад:
[tag = "V. *"] знаходить усі дієслова у всіх формах
[tag = "V.e. *"] шукає досконалих дієслів
[tag = "Va. *"] шукає всі екземпляри допоміжного дієслова "be"
[tag = "V..n. *"] шукає лише інфінітиви

Ви можете створювати запити, які шукають певну лему (всі морфологічні форми даного слова)
...
[lemma = "день"]
... або леми.
[lemma = "день | ніч | ранок | вечір"]

Поєднання атрибутів

Форма "мама" може належати до дієслова чи іменника. Щоб обмежити пошук однією з частин мови, потрібно додати атрибут "тег" з відповідним значенням у тому ж сегменті. Ми поєднуємо атрибути "word" та "tag", використовуючи символ & (ampersand).
[word = "мама" & tag = "V. *"]
Запит [lemma = "день" & tag = "N ... p. *"] Я знаходжу іменник день у формі множини.
Ви повинні стежити за:
використанням лапок під час пошуку значень;великими та малими літерами;великими та малими регістрами

Поєднання атрибутів

Форма "мама" може належати до дієслова чи іменника. Щоб обмежити пошук однією з частин мови, потрібно додати атрибут "тег" з відповідним значенням у тому ж сегменті. Ми поєднуємо атрибути "word" та "tag", використовуючи символ & (ampersand).

[word = "мама" & tag = "V. *"]

Запит [lemma = "день" & tag = "N ... p. *"] Я знаходжу іменник день у формі множини.

Ви повинні стежити за:

-використанням лапок під час пошуку значень

-великими та малими літерами

-великими та малими регістрами в тегах синтаксису

Які прийменки слідують за впливом?

[lemma = "вплив"] [tag = "PRP"]

Які прийменники слідують за впливом іменника?

[lemma = "вплив" & tag = "N .."] [tag = "PRP"]

Які прийменники дотримуються наступних синонімів?

[lemma = "боротьба | битва | боротьба"] [tag = "PRP"]

Як дозволити простір / включення між сегментами

Іноді доводиться розширювати варіанти пошуку, дозволяючи присутність інших слів між двома, які нас цікавлять.

Пара квадратних дужок без заливки означає будь-який сегмент.

[lemma = "день"] [] [lemma = "добре"]

Число між дужками {} після будь-якого відрізка, включаючи невизначений [], вказує на кількість слів, які повинні з’явитися між ними. Цей запит визначає включення будь-яких трьох слів між успіхом і успіхом.

[lemma = "звернутися"] [] {3} [lemma = "успіх"]

Використання {1,3} дає діапазон від 1 до 3, тобто щонайменше 1 і максимум 3 між опусканням і вниз.

[lemma = "хай"] [] {1,3} [word = "вниз"]

Як виключити предмет

Знак оклику перед знаком рівності не дорівнює. Наступний запит до корпусу BNC знайде швидко як іменник, дієслово, прислівник, але не як прикметник.

[lemma = "швидкий" & тег! = "AJ0"]

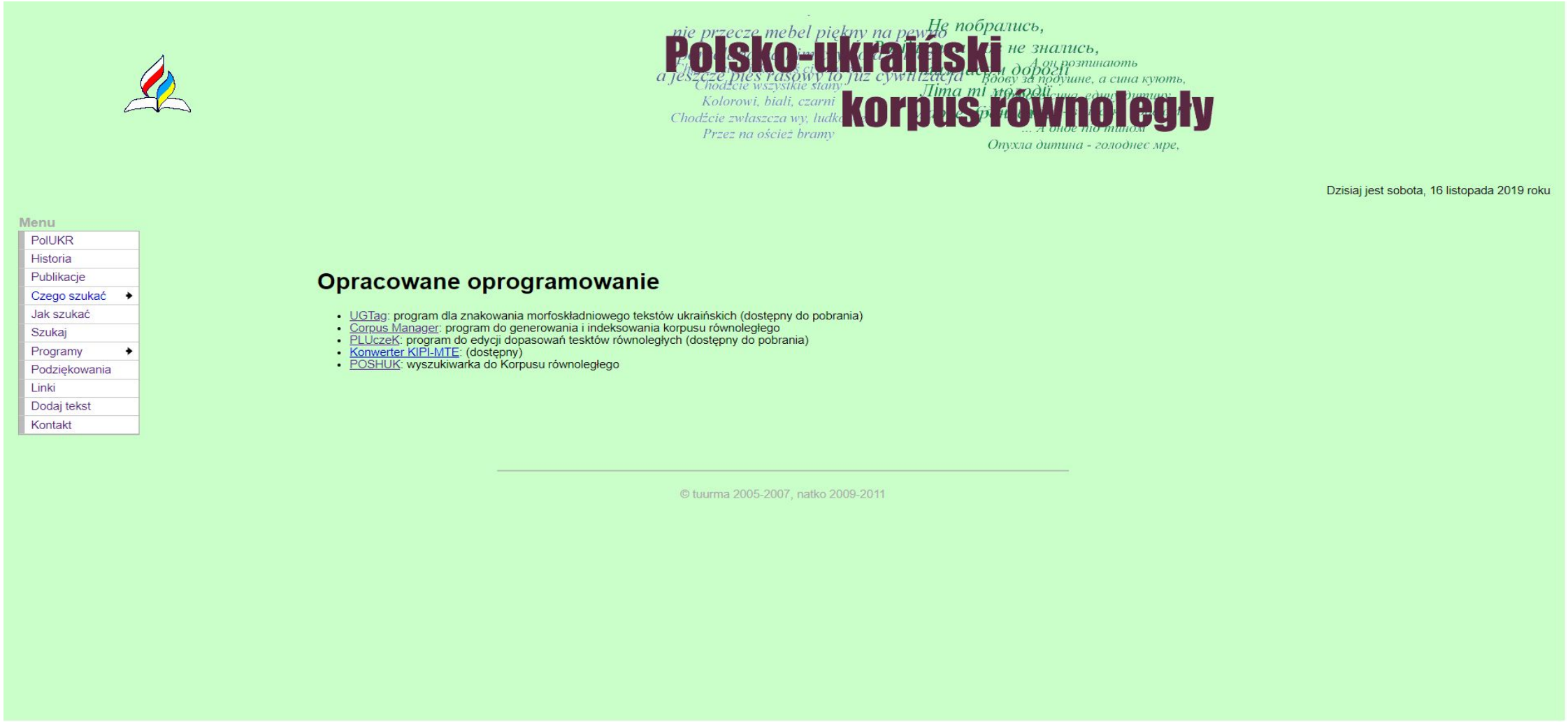
Наступний запит BNC шукає сновидіння, за яким випливає все, що завгодно.

[lemma = "мрія"] [слово! = "про"]

Наступні приклади знаходять усі форми перерви з наступними п’ятьма словами, а потім посміхаються не як дієслово.

[lemma = "перерва"] [] {5} [lemma = "усмішка" & тег! = "V .."]

ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ



Маркування

Тексти маркування включають структуру: глави, абзаци, речення, слова; метаінформація: автор, назва, перекладач (якщо це перекладений текст), рік та місце видання, жанр тощо. Тексти лематизовані, тобто кожен запис має свою задану форму вступу; крім того, вони містять розширену граматичну анотацію відповідно до рекомендованого міжнародного формату MULTEXT-East. Оригінальна граматична інформація для польської мови надходить з аналізатора Morpheus та TaKIPI, для української мови - із Словника української граматики та синтаксичного аналізатора морфа UGTAg. Заради узгодженості формату в PolUKR він був значно модифікований та розширений в обох випадках. Як польський, так і український набір граматичних тегів (англ. Tagset) у корпусі налічує понад 1200 унікальних граматичних кодів, які концептуально можна порівняти завдяки загальному формату.

Завдяки використанню міжнародних стандартів письма, а саме формату XML відповідно до рекомендацій TEI, а також охоплює найбільшу кількість мов серед існуючих граматичних форматів для написання MULTEXT-East, PolUKR має великий потенціал для розширення на подальші мови та інтеграції з існуючими мовними ресурсами. Це також єдиний загальнодоступний морфосинтезований орган української мови.

ДЯКУЮ ЗА УВАГУ!