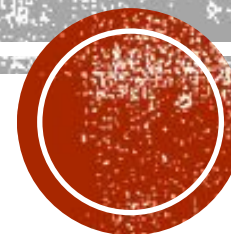


**POLSKO-UKRAIŃ
SKI KORPUS
RÓWNOLEGŁY**





Polsko-ukraiński korpus równoległy

Dzisiaj jest sobota, 16 listopada 2019 roku

Menu

PolUKR
Historia
Publikacje
Czego szukać ▾
Jak szukać
Szukaj
Programy ▾
Podziękowania
Linki
Dodaj tekst
Kontakt

Streszczenie projektu

Elektroniczny Polsko-Ukraiński Korpus Równoległy może być narzędziem do badań lingwistycznych w zakresie leksykologii, semantyki, gramatyki, stylistyki itd., a jednocześnie dużą bazą materiałową, wychodzącą naprzeciw potrzebom leksykografii polsko-ukraińskiej.

Teksty

Korpus zawiera teksty oryginalne i tłumaczone w językach polskim i ukraińskim, tworzone przeważnie w XX wieku, dopasowane (ang. *aligned*) na poziomie zdań i należące do różnych gatunków: literatura piękna, publicystyka, podręczniki, dokumenty, wiadomości prasowe, ogółem ok. 3 milionów wyrazów.

Znakowanie

Znakowanie tekstów obejmuje strukturę: rozdziały, akapity, zdania, wyrazy; meta informacje: autor, tytuł, tłumacz (jeżeli jest to tekst tłumaczony), rok i miejsce wydania, gatunek, itd. Teksty są lematyzowane, tzn. do każdej formy wyrazu została podana jej forma hasłowa; ponadto zawierają one rozszerzoną adnotację gramatyczną, zgodną z rekomendowanym międzynarodowym formatem MULTEXT-East. Oryginalna informacja gramatyczna dla języka polskiego pochodzi z analizatora Morfeusz i tagera TaKIP1, dla języka ukraińskiego – z Ukraińskiego Słownika Gramatycznego i analizatora morfoskładniowego UGTAg. Dla potrzeb spójności formatu w PolUKR została ona w obu przypadkach znacznie zmodyfikowana i rozszerzona. Zarówno polski, jak i ukraiński zestaw znaczników gramatycznych (ang. tagset) w Korpusie liczy ponad 1200 unikalnych gramatycznych kodów, które są porównywalne pojęciowo ze względu na wspólny format.

Wyszukiwarka

Korpus jest wyposażony w wyszukiwarkę POSHUK (skrót od wyrazów *POlski, UKraiński, Search*, cały wyraz po ukraińsku oznacza *wyszukiwanie*), która pozwala na tworzenie złożonych zapytań, łączących informacje o lemacie, formie gramatycznej, strukturze tekstu, meta informacji, itd. oraz stosowanie wyrażań regularnych. Dla wygody wyszukiwania zastosowany został także mechanizm aliasów dla tagów gramatycznych, np. bardziej intuicyjne *gen* zamiast *g* do określania dopełniacza czy *prep* zamiast *S* (przyimek). Warunki wyszukiwania można określać w obu częściach językowych Korpusu. Wyniki wyszukiwania w polskiej i ukraińskiej częściach można łączyć na różne sposoby: przekrój (AND), suma (OR) i różnica (XOR) wyników. Jest możliwe wyszukiwanie w obu językach albo tylko w jednym z języków, także tylko w tekstach oryginalnych bądź tylko tłumaczonych. Wyniki wyszukiwania są zapisywane w postaci plików html.

Perspektywy

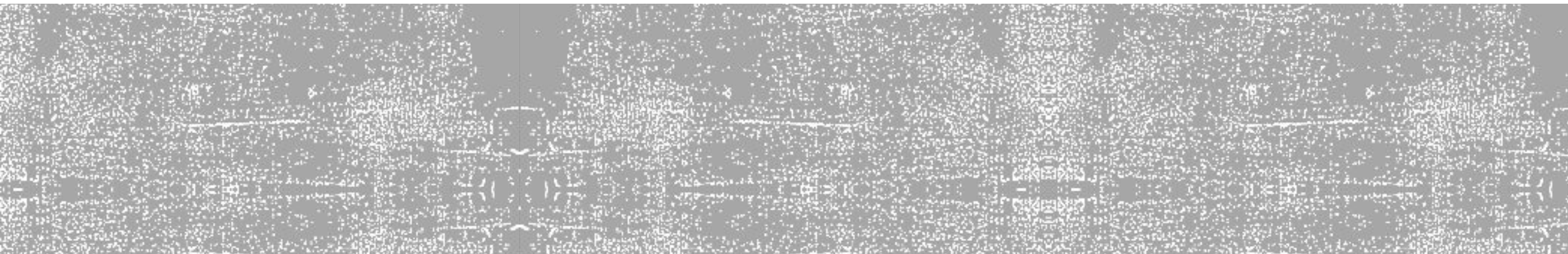
Ze względu na zastosowanie międzynarodowych standardów zapisu, mianowicie formatu XML zgodnego z rekomendacjami TEI, a także obejmującego największą liczbę języków wśród istniejących formatów zapisu gramatycznego MULTEXT-East, PolUKR ma duży potencjał do rozszerzenia na kolejne języki i integrację z istniejącymi zasobami językowymi. Jest on także jedynym dostępnym publicznie oznakowanym morfoskładniowo korpusem języka ukraińskiego.

© tuurma 2005-2007, natko 2009-2011

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.

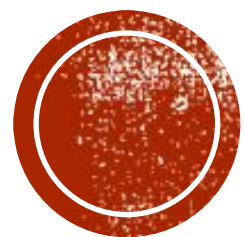
Тексти лематизовані та містять синтаксичну інформацію морфа, при цьому набори тегів для польської та української є стандартизованими.

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 <i>big bag beg bog bug</i>
gwiazdka *	po kropce znajduje dowolną liczbę dowolnych liter *iwać znajdzie wyrazy, które kończą się na <i>-iwać</i> przy.* znajdzie wyrazy, które zaczynają się na <i>przy-</i> bl.*ng znajdzie wyrazy, które zaczynają się na <i>bl-</i> i kończą się na <i>-ng</i> *sta.* znajdzie wyrazy, które mają w środku ciąg <i>sta</i> .
pytajnik ?	Znaki w nawiasach są opcjonalne. blond(e)? znajdzie <i>blond</i> oraz <i>blonde</i>
kreska pionowa	<i>walka bitwa wojna</i> szuka dowolnego z podanych wyrazów
nawiasy kwadratowe [i]	Mogą zawierać znaki pojedyncze alternatywne. «[r]pati» znajdzie dwa warianty pisowni « <i>grati</i> » i « <i>ḡrati</i> »

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 / і 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 (аббревіатура, аббревіатура), X (невизнаний, залишковий).

Див. детальний опис для польської та української

Синтаксис запити:

```
[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: р недосконалий (прогресивний) та досконалий. Четверте місце займає інформація про форму дієслова: і - вказівний режим, с - 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 ... р. *"] 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 ... р. *"] 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 слідуєть за впливом?

```
[lemma = "вплив"] [tag = "PRP"]
```

Які прийменники слідуєть 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 впливає 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



nie przecze, mebel piękny na pewno. Не побрались,
a jeszcze pies rasowy to już cywilizacja. не знались,
Chodźcie wszystkie stany. А ви розминають
Kolorowi, biali, czarni. Воду за подушце, а сина кують.
Chodźcie zwłaszcza wy, ludko. Тима ти могоді.
Przez na oścież bramy. ... А бібе під тилом
Опухла дитина - голоднес мре.

Dzisiaj jest sobota, 16 listopada 2019 roku

Menu

PolUKR
Historia
Publikacje
Czego szukać →
Jak szukać
Szukaj
Programy →
Podziękowania
Linki
Dodaj tekst
Kontakt

Opracowane oprogramowanie

- [UGTag](#): program dla znakowania morfokładniowego tekstów ukraińskich (dostępny do pobrania)
- [Corpus Manager](#): program do generowania i indeksowania korpusu równoległego
- [PLUczek](#): program do edycji dopasowań tekstów równoległych (dostępny do pobrania)
- [Konwerter KIPJ-MTE](#): (dostępny)
- [POSHUK](#): wyszukiwarka do Korpusu równoległego



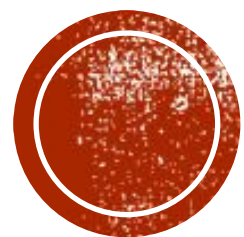
Маркування

Тексти маркування включають структуру: глави, абзаци, речення, слова; метайнформація: 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!