Граматичний потенціал Польсько-українського паралельного корпусу (PolUKR)

Наталя Коциба
Інститут Міждисципланірних Досліджень, Варшавський університет,
20 листопада 2009 р.

План презентації

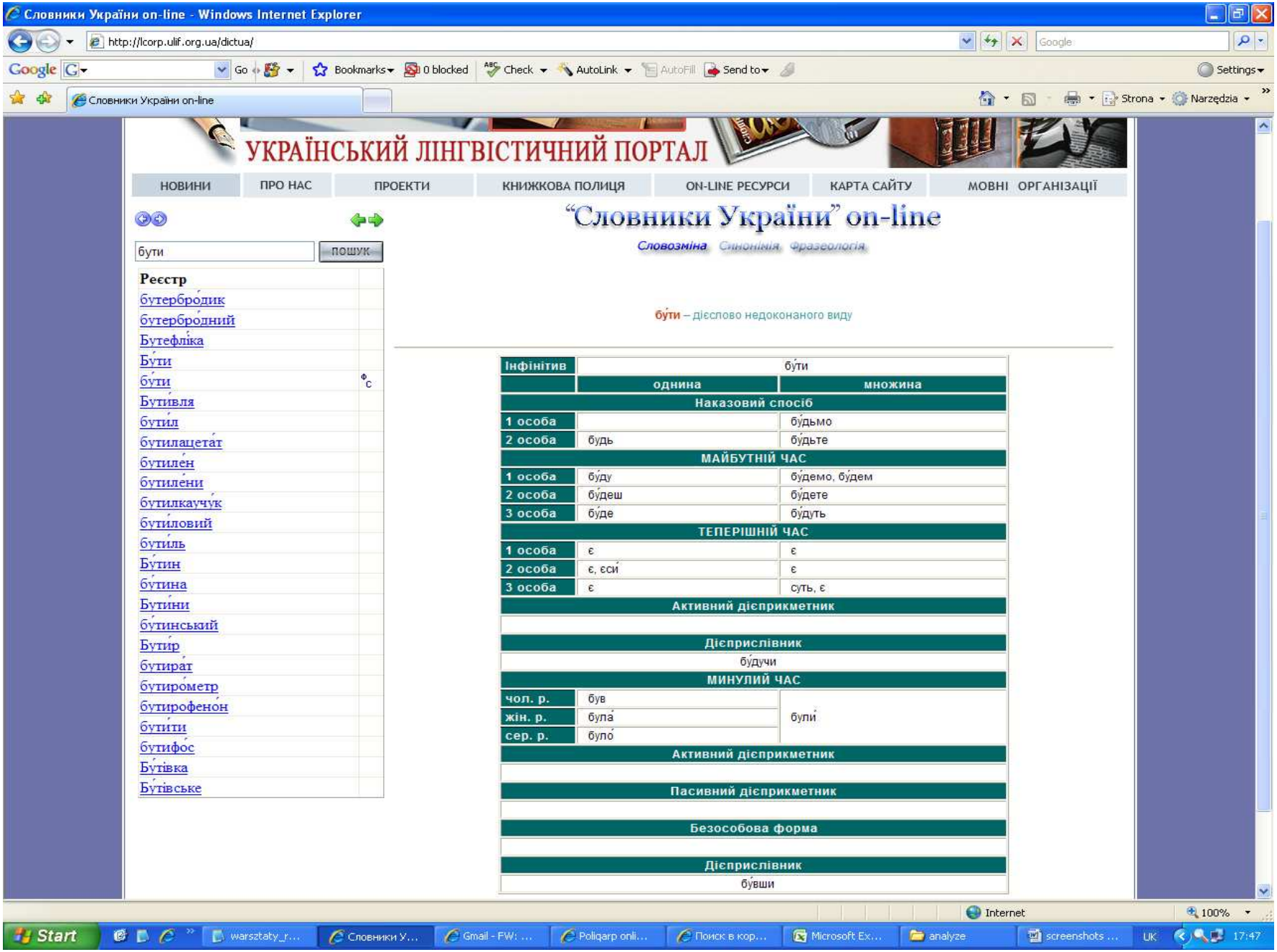
- історія створення PolUKRy, можливості його застосування в лексикографії: http://www.domeczek.pl/~polukr
- *два типи тагсетів в PolUKR: розширення тагсету Корпусу IPI PAN і MULTEXT-East*
- демонстрація програми УГТаг і пошуку по корпусу (Андрій Микуляк, Інститут Ядерної Фізики, Варшава)
- етапи роботи над корпусом: від паперових книжок до формату XML; демонстрація програми PLUczeK

Морфосинтактичне тагування в PolUKR

- джерела морфосинтактичної інформації в PolUKR
- види таґсетів залежно від способу запису граматичної інформації (форма представлення)
- вибрані „проблематичні” категорії і як їх трактують існуючі таґсети (зміст)

Джерела морфосинтактичної інформації

- Польська мова: пакет TaKIPI, таґер Корпусу ІПІ (ПАН) + Вроцлавська політехніка, біля 200 тис. слів
- лематизатор, морфологічний аналізатор, уоднозначнювач
- Українська мова: граматичний словник УМІФ НАНУ, 250 тис. слів
- лематизатор, морфологічний аналізатор, немає уоднозначнення


Словники України on-line - Windows Internet Explorer
http://lcorp.ulif.org.ua/dictua/
Словники України on-line
УКРАЇНСЬКИЙ ЛІНГВІСТИЧНИЙ ПОРТАЛ
НОВИНИ
ПРО НАС
ПРОЕКТИ
КНИЖКОВА ПОЛИЦЯ
ON-LINE РЕСУРСИ
КАРТА САЙТУ
МОВНІ ОРГАНІЗАЦІЇ
бути
ПОШУК
Реєстр
бутербро́дик
бутербро́дний
Бутефлі́ка
Бу́ти
бу́ти
Бутивля
бути́л
бутилацета́т
бутиле́н
бутилени
бутилкаучу́к
бути́ловий
бути́ль
Бу́тин
бу́тина
Бути́ни
бу́тинський
Бути́р
бутира́т
бутиро́метр
бутирофено́н
бути́ти
бутифо́с
Бу́тівка
Бу́тівське
“Словники України” on-line
Словозміна Синонімія Фразеологія
бу́ти – дієслово недоконаного виду
Інфінітив
бу́ти
однина
множина
Наказовий спосіб
1 особа
бу́дьмо
2 особа
будь
бу́дьте
МАЙБУТНІЙ ЧАС
1 особа
бу́ду
бу́демо, бу́дем
2 особа
бу́деш
бу́дете
3 особа
бу́де
бу́дуть
ТЕПЕРІШНІЙ ЧАС
1 особа
є
є
2 особа
є, єси́
є
3 особа
є
суть, є
Активний дієприкметник
Дієприслівник
бу́дучи
МИНУЛИЙ ЧАС
чол. р.
був
були́
жін. р.
була́
сер. р.
було́
Активний дієприкметник
Пасивний дієприкметник
Безособова форма
Дієприслівник
бу́вши
Internet
100%
Start
17:47

Український нетаґований текст

- Львів розташований на етнічних українських землях і є одним з головних нервових вузлів українського народу, найважливішим клапаном його серця, вічним збудником честолюбства, гордості й потягу до волі.

Українська: вхідні таґи з лемами

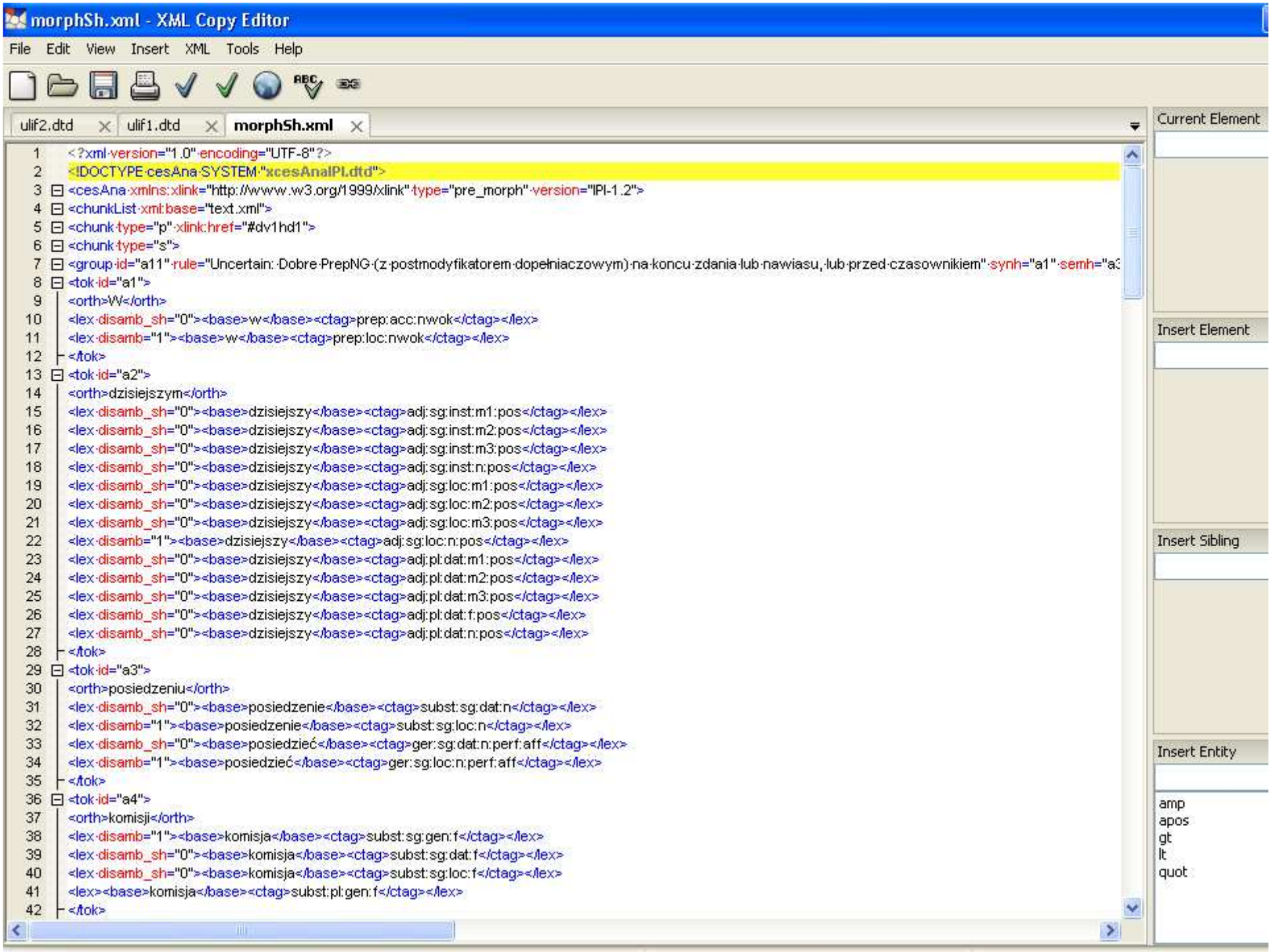
Львів<JDJAJIJK><Львів 0|Львів 0|Лев 1|Лев 1|>
розташований<BDBAV?><розташований 0|розташований 0|розташувати 0|>
на<N0N0Z0PF><на 4|на 3|на 2|на 1|> **етнічних**<AVATAX><етнічний 0|етнічний 0|етнічний 0|> **українських**<AVATAXJIGIJKGKJMGM><український 0|український 0|український 0|Український 0|Український 0|>
землях<FM><земля 2|> **і**<SSSCN0Z0><і 1|і 3|і 2|> **є**<UPUOUNUKUMUL><бути 0|бути 0|бути 0|бути 0|бути 0|бути 0|> **одним**<HUHQHERQRERU><один 0|один 0|один 0|оден 0|оден 0|оден 0|> **з**<PE><з 0|> **головних**<AVATAX><головний 0|головний 0|головний 0|> **нервових**<AVATAX><нервовий 0|нервовий 0|нервовий 0|> **вузлів**<MIMI><вузол 2|вузол 1|>
українського<ANADABJDJBKB><український 0|український 0|український 0|Український 0|Український 0|Українське 0|> **народу**<MBMCMBMC><народ 0|народ 0|нарід 0|нарід 0|>, **найважливішим**<AQAEAU><найважливіший 0|найважливіший 0|найважливіший 0|> **клапаном**<ME><клапан 0|>
його<FGODOBODOB><його 0|воно 0|воно 0|він 0|він 0|>
серця<NKNHNBNN><серце 0|серце 0|серце 0|серце 0|>,
вічним<AQAEAU><вічний 0|вічний 0|вічний 0|> **збудником**<MEME><збудник 1|збудник 2|> **честолюбства**<NB><честолюбство 0|>,
гордості<FCFBFF><гордість 0|гордість 0|гордість 0|> **й**<SSSCZ0><й 1|й 2|>
потягу<MFMCMGMBMCMFMGFDGD><потяг 2|потяг 2|потяг 2|потяг 1|потяг 1|потяг 1|потяг 1|потяга 0|Потяга 0|>
до<NGNFNENDNCNBNANHNINJNKNLNMNNPB><до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 1|>
волі<UOFCFBFFGCGBGFGFGCGBGHGNNFMF><воліти 0|воля 0|воля 0|воля 0|Воля 2|Воля 2|Воля 2|Воля 1|Воля 1|Воля 1|Воля 1|Воля 1|воло 0|віл 0|>.

Приклади граматичних кодів, що використовуються в УМІФ (384)

Граматичне значення	**Морфкод**	**Приклад**
Дієслово, інфінітив, доконаний вид, активний стан	VA	*прочитати*
Дієприкметник, чоловічий рід, однина, називний відмінок, доконаний вид, минулий час, активний	BA	*зрослий*
Невідмінюваний прикметник	AZ	*ультра*
Іменник загальний, жіночий рід, однина, давальний відмінок	FC	*квітці*
Предикатив (присудкове слово)	X0	*слід*

Польський нетаґований текст

- W dzisiejszym posiedzeniu komisji uczestniczy ekspert komisji pan profesor Jan Gajewski.


morphSh.xml - XML Copy Editor
File Edit View Insert XML Tools Help
ulif2.dtd
ulif1.dtd
morphSh.xml
Current Element
Insert Element
Insert Sibling
Insert Entity
amp
apos
gt
lt
quot

Види таґсетів залежно від способу запису граматичної інформації (форма презентації)

Види таґсетів залежно від рівня кодування (словоформа чи значення притаманних їй категорій)

СИМВОЛЬНІ: граматична характеристика словоформи передається в одному коді British National Corpus, корпус УМІФ

вимагає мало машинної пам'яті, але занадто багато людської

ЛАНЦЮЖКОВІ: кожна категорія, її атрибут і вартості цього атрибута мають свій унікальний код; морфосинтактична характеристика словоформи представлена послідовністю кодів вартостей атрибутів KIPI, CzNK, НКРЯ

можуть бути економніші при пошуку, ніж символи, напр. коли шукаємо допасування на рівні вартості якогось атрибуту, що властивий кільком частинам мови (відмінок, рід для іменника і прикметника)

Види таґсетів відносно позиціювання граматичної інформації

- **ПОЗИЦІЙНІ** CzNC, MULTEXT

Кожна категорія (виражена її вартістю в тазі) має фіксовану позицію у ланцюжку, вартості різних категорій можуть мати ту саму нотацію (t,n,y,p, 1, 2, 3)

ЧНК (біля 4 тис.)

volen: VsYS---XX-AP--- verb, passive participle, masculine, singular, any person, any tense, positive, passive

hraniční: AAIS4----1A---- standard adjective, masc. inanimate, singular, accusative, positive

Щоб задати пошук по одному з атрибутів, треба записати цілий ланцюжок

- ***УНІВЕРСАЛЬНІ** KIPI „.*gen.*”, „case=„gen””
- (символьних кодів ця проблема не торкається)

Види таґсетів залежно від принципів та глибини категоризації граматичної інформації

- **ГНІЗДОВІ** (CzNC : 1.POS - 10 і 2.subPOS - 75)

- **ФЛЕКСЕМНІ** корпус ІПІ ПАН (польська)

Флексеми :: ЧМ (Януш Бєнь) множини слів зі спільними наборами змінних граматичних атрибутів

- Частини мови або їх підмножини об”єднуються у спільні класи флексем, якщо мають однакову граматичну характеристику (теоретична підстава до розбиття займенників на групи, що приєднуються до класів слів, які вони “заміняють”; виокремлення займенника *sie,siebie,winien*) **29**
- Принцип аглютинації для економності запису

Граматичні категорії польських флексем і їх вартості

- **number**: sg, pl;
- **case**: nom, acc, gen, dat, inst, loc, voc;
- **gender**: masculine personal m1(*facet*), masculine animate m2 (*koń*), masculine inanimate m3 (*stół*), feminine f (*kobieta, żyrafa, książka*), два середні роди: n1 (*dziecko*), n2 (*okno*), і три *plurale tantum* p1 (*wujostwo*), p2 (*drzwi*), p3 (*okulary*);
- **person**: pri, sec, ter;
- **degree**: pos, comp, sup;
- **aspect**: imperf, perf;
- **negation**: aff, neg;
- **accentability** (Pol.: *akcentowość*): akc, nakc;
- **post-prepositionality** (Pol.: *poprzyimkowość*): praep, npraep;
- **accommodability** (Pol.: *akomodacyjność*): congr, rec;
- **agglutination** (Pol.: *aglutynacyjność*): nagl, agl;
- **vocability** (Pol.: *wokaliczność*): wok, nwok.

Механізм аліасів

n	n1 n2
P	p1 p2 p3
masc	m1 m2 m3
noun	subst depr ger xxs ppron12 ppron3 PPRON GNOUN PROPNOUN
pron	ppron12 ppron3 siebie PPRON
verb	fin praet aglt będzie inf imps impt **pact ppas pcon pant** ger winien PART(PPAST) FUT PRES

Інтерпретація вибраних категорій в граматичних словниках і корпусах

Дієприслівники

- польська (вид вписано на рівні словника і визначає синхронність чи попередність, дієприслівники часу не мають, час дієслів визначається за допомогою виду або аналітично)

Флексеми:

adv.pres.prtcp.

adv.anter.prtcp.

- українська (вид і час – морфологічна характеристика)

1. Дієслово, дієприслівник, доконаний вид, *минулий час, активний стан **VW** *прочитавши* (піду, прочитавши)
2. Дієслово, дієприслівник, недоконаний вид, минулий час, активний стан **UW** *читавши*.
3. Дієслово, дієприслівник, недоконаний вид, *теперішній час, активний стан **UQ** *читаючи* (робив/робитиме читаючи)

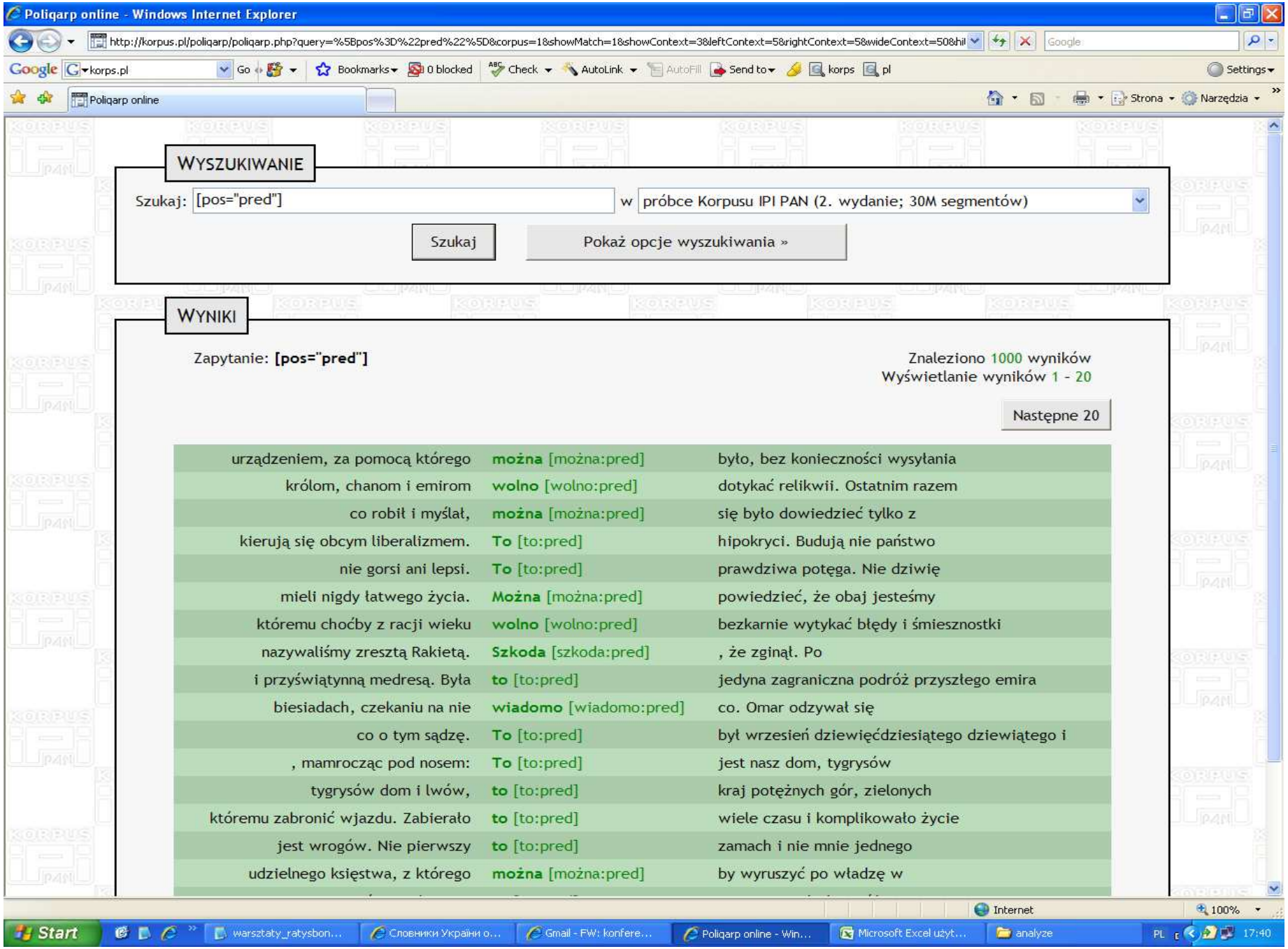
- PolUKR

adverbial and adjectival participles characterised by aspect and *tense:

verb:part:perf , verb:part:imperf, verb:part:imperf:praet

Займенники

- проблема слов'янських займенників: 296 таґів для 309 займенників (Elena Paskaleva)
- польська: поділ на 1-2 ос., 3 ос і *siebie* (*ów, jak?*)
- українська: займенник-іменник, займенник-прикметник
- російська: також займенник-предикатив і займенник-прислівник
- чеська: 18 підкатегорій на рівні підЧМ
- PolUKR: підхід УМІФ, додатково поділ ІПІ ПАН на 1-2 і 3 особові


Poliqarp online - Windows Internet Explorer
http://korpus.pl/poliqarp/poliqarp.php?query=%5Bpos%3D%22pred%22%5D&corpus=1&showMatch=1&showContext=3&leftContext=5&rightContext=5&wideContext=50&hil
Poliqarp online
WYSZUKIWANIE
Szukaj: [pos="pred"]
w próbce Korpusu IPI PAN (2. wydanie; 30M segmentów)
Szukaj
Pokaż opcje wyszukiwania »
WYNIKI
Zapytanie: [pos="pred"]
Znaleziono 1000 wyników
Wyświetlanie wyników 1 - 20
Następne 20
urządzeniem, za pomocą którego można [można:pred] było, bez konieczności wysyłania
królom, chanom i emirom wolno [wolno:pred] dotykać relikwii. Ostatnim razem
co robił i myślał, można [można:pred] się było dowiedzieć tylko z
kierują się obcym liberalizmem. To [to:pred] hipokryci. Budują nie państwo
nie gorsi ani lepsi. To [to:pred] prawdziwa potęga. Nie dziwię
mieli nigdy łatwego życia. Można [można:pred] powiedzieć, że obaj jesteśmy
któremu choćby z racji wieku wolno [wolno:pred] bezkarnie wytykać błędy i śmieszności
nazywaliśmy zresztą Rakietą. Szkoda [szkoda:pred] , że zginął. Po
i przyświątynną medresą. Była to [to:pred] jedyna zagraniczna podróż przyszłego emira
biesiadach, czekaniu na nie wiadomo [wiadomo:pred] co. Omar odzywał się
co o tym sądzę. To [to:pred] był wrzesień dziewięćdziesiątego dziewiątego i
, mamrocząc pod nosem: To [to:pred] jest nasz dom, tygrysów
tygrysów dom i lwów, to [to:pred] kraj potężnych gór, zielonych
któremu zabronić wjazdu. Zabierało to [to:pred] wiele czasu i komplikowało życie
jest wrogów. Nie pierwszy to [to:pred] zamach i nie mnie jednego
udzielnego księstwa, z którego można [można:pred] by wyruszyć po władzę w
Internet
100%
Start
warsztaty_ratysbon...
Словники України о...
Gmail - FW: konfere...
Poliqarp online - Win...
Microsoft Excel użyt...
analyze
PL
17:40

Предикативи польські

польська 26, українська 176, російська >1200, чеська 0

To ksiazka (This <is a> book) → займенник-іменник

Модальні слова: *można, trzeba, wolno, wiadomo, trza, niepodobna, podobna, dość, dosyć* →модальні прислівники (диференціація на семантичному рівні).

Віддієслівні: *słychać, widać, stać, czuć, znać* і відіменникові: *szkoda, potrzeba, żal, wstyd, strach, pora, czas, brak, śmiech* → прислівники

Предикативи укр.

- семельфактиви: *зирк, круть-верть* → вигуки
- відприслівникові: *зимно, безвітряно, безсніжно* → прислівники
- демінутивні дієслова: *спунькати, спатки, ходитоньки, їсточки, їстки* → інфінітивні форми дієслова

Прикметники

- УГС УМІФ: прикметники вищого і найвищого ступенів подано як окремі леми, ступінь як категорія не числиться взагалі
- IPI PAN:po-polsku;(adj vs по-польськи adv) polsko-niemiecki ???
- НКРЯ вищий ступінь прикметників і прислівників включено до предикативів
- PolUKR: розроблено правила визначення ступеня прикметника і його релематизації, реалізація на рівні конвертації таґів; післяприйменникові прикметники трактуються як частини прислівників

Таґсет PolUKR

- позиційний
- гніздовий (ЧМ і підЧМ) для користувача
- альяси підключаються на рівні пошуку

Фрагмент українського тексту з тагами PolUKR

morph_utf - WordPad

Plik Edycja Widok Wstaw Format Pomoc

Courier New 10 Cyrylica

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE cesAna SYSTEM "xcesAnaIPI.dtd">
<?xml-stylesheet href="display.css" type="text/css"?>
<csAna>
<chunkList>
<tok><orth>Опубліковані</orth>
<lex><base>Опубліковані</base><ctag>adj:pl:nom</ctag></lex>
<lex><base>Опубліковані</base><ctag>adj:pl:acc</ctag></lex>
</tok>
<tok><orth>наприкінці</orth>
<lex><base>наприкінці</base><ctag>adv</ctag></lex>
</tok>
<tok><orth>минулого</orth>
<lex><base>минулого</base><ctag>adj:sg:gen:masc</ctag></lex>
<lex><base>минулого</base><ctag>adj:sg:acc:masc</ctag></lex>
<lex><base>минулого</base><ctag>adj:sg:gen:n</ctag></lex>
</tok>
<tok><orth>тижня</orth>
<lex><base>тижня</base><ctag>gnoun:sg:gen:masc</ctag></lex>
</tok>
<tok><orth>у</orth>
<lex><base>у</base><ctag>int</ctag></lex>
<lex><base>у</base><ctag>prep</ctag></lex>
<lex><base>у</base><ctag>prep</ctag></lex>
<lex><base>у</base><ctag>prep</ctag></lex>
</tok>
<tok><orth>Financial</orth>
<lex><base>Financial</base><ctag>ign</ctag></lex>
</tok>
<tok><orth>Time</orth>
<lex><base>Time</base><ctag>ign</ctag></lex>
</tok>
<tok><orth>звинувачення</orth>
<lex><base>звинувачення</base><ctag>gnoun:sg:nom:n</ctag></lex>
<lex><base>звинувачення</base><ctag>gnoun:sg:gen:n</ctag></lex>
<lex><base>звинувачення</base><ctag>gnoun:sg:acc:n</ctag></lex>
<lex><base>звинувачення</base><ctag>gnoun:sg:voc:n</ctag></lex>
<lex><base>звинувачення</base><ctag>gnoun:pl:nom:n</ctag></lex>
```

Aby uzyskać Pomoc, naciśnij klawisz F1. NUM

2008 Zakopane | KINGSTON (G:) | Microsoft PowerPoi... | morph_utf - WordPad | PL 19:34

Тагсети MTE

- Міжнародний проект MULTEXT-East: http://nl.ijs.si/ME/
- морфосинтаксичні специфікації для польської і української мов від версії MTE-4
- конвертер для польської мови з формату Корпусу IPI PAN (скор. КІПІ) до MTE-4(pl), доступний за адресою: http://domeczek.pl/~polukr/mte-conv

Конверсія польських тагів

- Конверсію зроблено на основі формату тагів морфологічного аналізатора Морфеуш, які також використовуються для Корпусу IPI PAN. З розмічених за допомогою програми TaKIPI текстів корпусу витягнуто автоматично1295 тагів.
- Потім таги були поділені на їх найменші значеннєві частини і записані до реляційної бази даних, де кожна вартість займала окрему колонку. Коди були замінені на їх відповідники в MTE
- Вигенеровано нові коди з відповідним порядком атрибутів для кожної частини мови.
- Велика частина тагів мала однозначну проекцію, але багато з них також були поділені на групи.

Типи проекцій тагів

- Однозначна проекція на MTE (1192 тагів): порівняльний і найвищий ступінь прикметників, дієслова, дієприкметники, герундії (відієслівні іменники), кількісні іменники, депреціативні іменники, особові і зворотні займенники, множина іменників, прийменники.
- Таги, котрі поділено на наступні категорії в MTE: насамперед “кублики” і неособові займенники.
- Нові таги: збірні числівники, кілька форм займенників.
- Таги, котрі були зведені до одного.

Розширення тагів КІПІ

- З 1298 оригінальних тагів 101 отримали більше, ніж одну проекцію в MTE:
- кожен з 60 тагів для прикметників основного ступеня порівняння був поділений на 13 більш детальних;
- 18 іменникових тагів “subst” отримали від 2 до 7 тагів в MTE;
- “кублики” були поділені на 7 категорій з 27 унікальними тагами
- предикативи поділено на 3 категорії з 4 тагами

Розроділ “кубликів” в проекції MTE

Категорія	Приклад	Таг MTE	Кількість лексем
C	albo ż	1	11
I	hej	1	179
P	jakoś, się	16	85
Q	że	2	74
R	wczoraj	4	233
S	ponad	2	7
X	mocium	1	8

Нові таги

Таг КІПІ	Таг MTE	Опис тагу MTE	Слововживання	Приклад
ppron3:sg:gen:f:ter:nakc:praep	Pp-3f--sgy-n	Pronoun Type=personal Person=third Gender=feminine Number=singular Case=genitive Clitic=yes Syntactic_Type=nominal	44	*niej*
ppron3:sg:gen:f:ter:nakc:praep	Pp-3f--sgasn	Pronoun Type=personal Person=third Gender=feminine Number=singular Case=genitive Clitic=agglutinant Definiteness=short-art Syntactic_Type=nominal		*ń*
ppron3:sg:acc:f:ter:nakc:praep	Pp-3f--say-n	Pronoun Type=personal Person=third Gender=feminine Number=singular Case=accusative Clitic=yes Syntactic_Type=nominal	11	*nią*
ppron3:sg:acc:f:ter:nakc:praep	Pp-3f--saasn	Pronoun Type=personal Person=third Gender=feminine Number=singular Case=accusative Clitic=agglutinant Definiteness=short-art Syntactic_Type=nominal		*ń*

Скорочення тагів КІПІ, статистика

- особові займенники 3-тьої особи (флексема ppron3 у КІПІ) передбачають 287 різних тагів КІПІ для опису 5 лем з їх 23 формами → 65 тагів MTE.
- 1-ша і 2-га особа (флексема ppron12); 146 оригінальних тагів КІПІ → 30 тагів MTE.
- 433 таги для 42 форм особових займенників у КІПІ зведено до 95 тагів у версії MTE
- таги на словоформу: починаючи від форми *nim* з 53 інтерпретаціями в КІПІ, 33 для *nich* і 25 для *nimi* (16 форм з 10 чи більше інтерпретаціями) до *mu*, *jemu*, *ją* з 3 чи 4 інтерпретаціями.

Таги особових займенників жін. роду третьої особи однини у знахідному відмінку

Таг КІПІ	Таг MTE	Форма слова
ppron3:sg:acc:f:ter:akc:npraep	Pp-3f--san-n	*ją*
ppron3:sg:acc:f:ter:akc:praep	Pp-3f--say-n	*nią*
ppron3:sg:acc:f:ter:nakc:npraep	Pp-3f--san-n	*ją*
ppron3:sg:acc:f:ter:nakc:praep	Pp-3f--say-n	*nią*
ppron3:sg:acc:f:ter:npraep	Pp-3f--san-n	*ją*
ppron3:sg:acc:f:ter:praep	Pp-3f--say-n	*nią*

Інтерпретація:
Pp-3f--san-n: Pronoun Type=personal Person=third Gender=feminine Number=singular Case=accusative Clitic=no Syntactic_Type=nominal
Pp-3f--say-n: Pronoun Type=personal Person=third Gender=feminine Number=singular Case=accusative Clitic=yes Syntactic_Type=nominal

Поділ на слова у реченнях: *moglibyście*

КІПІ

- <orth>mogli</orth><lex disamb="1"><base>móc</base><ctag>praet:pl:m1:imperf</ctag></lex><ns/>
- <orth>by</orth><lex disamb="1"><base>by</base><ctag>qub</ctag></lex><ns/>
- <orth>ście</orth><lex disamb="1"><base>być</base><ctag>aglt:pl:sec:imperf:nwok</ctag></lex>

MTE

- <w lemma="móc" ana="Vmpis-pmy">mogli</w>
  <w lemma="by" ana="Q">by</w>
  <w lemma="być" ana="Vapip2p--sa">ście</w>

Переглянутий поділ на слова:

- <w lemma="móc" ana="Vmpis2pmy-y">mogliście</w>
- <w lemma="móc" ana="Vmpcp3pmy-y">mogliby</w>
- <w lemma="móc" ana="Vmpcp2pmy-y">moglibyście</w>

Фрагмент індексу MSD

Таг MTE	Опис тагу	Лексеми	Приклад
Vmeis2sf--y	Verb Type=main Aspect=perfective VForm=indicative Tense=past Person=second Number=singular Gender=feminine Clitic=yes	85	*powiedziałaś/powiedzieć, zrobiłaś/zrobić, przyszłaś/przyjść*
Vmeis2sm--y	Verb Type=main Aspect=perfective VForm=indicative Tense=past Person=second Number=singular Gender=masculine Clitic=yes	274	*przyszedłeś/przyjść, powiedziałeś/powiedzieć, zrobiłeś/zrobić,*
Vmeis2sn--y	Verb Type=main Aspect=perfective VForm=indicative Tense=past Person=second Number=singular Gender=neuter Clitic=yes	1	*pozostałoś/pozostać, przeszłoś/przejść*
Vmeis-pf	Verb Type=main Aspect=perfective VForm=indicative Tense=past Person=second Number=plural	619	*odbyły/odbyć, rozpoczęły/rozpocząć, zaszły/zajść*

Фрагмент лексикону

- absurdami absurd N-mnnpi 17
- absurdem absurd N-mnnsi 307
- absurdom absurd N-mnnpd 6
- absurdowi absurd N-mnnsd 4
- absurdu absurd N-mnnsg 578
- absurdy absurd N-mnnpa 59
- absurdy absurd N-mnnpn 58
- absurdzie absurd N-mnnsl 17
- absurdów absurd N-mnnpg 163
- aby aby C 201168
- ac ac X 1099
- ach ach I 1170

15 тисяч найбільш вживаних лем витягнуто за допомогою програми Poliqarp

У лексиконі налічується 175848 форм цих слів (приблизно 11.72 форм на лему).

Фрагмент тагованого тексту польського перекладу Джорджа Оруела “*1984*”

- <p id="Opl.5">
- <s id="Opl.5.1">
- <w lemma="być" ana="Vmpis-sm">Był</w>
- <w lemma="jasny" ana="A-pm--sn">jasny</w>
- <c>,</c>
- <w lemma="zimny" ana="A-pm--sn">zimny</w>
- <w lemma="dzień" ana="N-mnnsa">dzień</w>
- <w lemma="kwietniowy" ana="A-pmn-sa">kwietniowy</w>
- <w lemma="i" ana="C">i</w>
- <w lemma="zegar" ana="N-mnnpn">zegary</w>
- <w lemma="bić" ana="Vmpis-pmn">biły</w>
- <w lemma="trzynasty" ana="Mlof--si">trzynastą</w>
- <c>.</c>
- </s>

Бібліографія

- INTERA unified tagset project www.elda.org/intera
- Hanna Dalewska-Greń *Języki Słowiańskie, Warszawa,* PWN 1997.
- Tomas Erjavec et al. *Multext-East specifications for Slavic languages*, Budapest, 2003.
- Jan Hajič. Positional Tags: Quick Reference (Czech „HM" Morphology), 2000.
- Elena Paskaleva. *Balcan South-East Corpora Aligned to English*. In: *The Proceedings of the Workshop on Common Natural Language Processing Paradigm for Balkan Languages, EACL 2007*
- Adam Przepiórkowski and Marcin Woliński. *A Flexemic Tagset for Polish*. In: *The Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003.* http://nlp.ipipan.waw.pl/~adamp/Papers/2003-eacl-ws12/ws12.pdf
- Широков В.А et al. *Корпусна лінгвістика*. Київ: Довіра, 2005.

Документація до PolUKR

статті і слайди доступні зі сторінки http://www.domeczek.pl/~natko/

The Current State of Work on the Polish-Ukrainian Parallel Corpus (PolUKR). Proceedings of the International Workshop within MONDILEX project "Problems of Slavic Lexicography" Kyiv, 2-4 February 2009.

Морфосинтаксичне тагування польсько-українського паралельного корпусу (PolUKR). Proceedings of the International Conference “MegaLing'2008. Horizons of Applied Linguistics and Linguistic Technologies, Parthenit – Crimea, Ukraine, 20-27 September 2008”.

Ivan Derzhanski and Natalia Kotsyba. Towards a Consistent Morphological Tagset for Slavic Languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian. Metalanguage and Encoding Scheme Design for Digital Lexicography: MONDILEX Third Open Workshop, Bratislava, 15–16 April 2009.

Natalia Kotsyba, Andriy Mykulyak, Igor V. Shevchenko. UGTag: morphological analyzer and tagger for Ukrainian language. Proceedings of the international conference Practical Applications in Language and Computers (PALC 2009), Łódź, 6-8 April 2009.

Natalia Kotsyba, Adam Radziszewski and Ivan Derzhanski. Integrating the Polish language into the MULTEXT-East family: morphosyntactic specifications, converter, lexicon and corpus. (presentation slides). Proceedings of Research Infrastructure for Digital Lexicography: MONDILEX Fifth Open Workshop, October 14, 2009, Ljubljana, Slovenia, within INFORMATION SOCIETY 2009 12th International multiconference.