Korpus języka polskiej społeczności na Ukrainie i metody jego utworzenia

Aleksandra Wieczorek

Warszawa, 20 lutego 2012

- nagrania z polskiej wsi Maćkowce na Ukrainie, przepisane ręcznie – zapis fonetyczny
- korpus lematyzowany, zachowany zapis fonetyczny
- ok. 140 tys. słów informatorów
  - NKRJaDial – 250 tys. słów (2008), anotowany, zapis częściowo standaryzowany
  - czeskie korpusy mówione – 500 tys.– 1 mln słów, nieanotowane
- program FonOrt (Michał Wieczorek)

Uproszczony zapis fonetyczny

- ż [ž]
- sz [š]
- cz [č]
- .e [ė]
- .o [ȯ]
- & [ə]
- u([ų̂]
- u$ [ų]

Przykład

[*Rodzice uczyli panią mówić po polsku?*] N… fszyscy pu polsku, w'oska… pu(`ołnośću polska. Maćk… Maćk$^{u(}$ofc.e, Szaraw'eczka, Ryczan.e i Zażecz.e. Sztyry is śuł u nas u tutej polsk'ich, i my pu polsku ruzmaw'amy. Du wojny była polska szkoła, ja p'irsze klase kuńczyła polske. Bukf`ar był ftedy i `azbuke my uczyli.

Etapy budowy korpusu

- konwersja do postaci XML
- interpretacja zapisu
- lematyzacja
- przypisanie kwalifikatora etymologicznego oraz znaczenia
- przystosowanie do wyszukiwarki Poliqarp

Konwersja do postaci XML

- zachowanie oznaczeń osoby mówiącej, nazw własnych, tematu rozmowy itp.

- np.

  [*Rodzice uczyli panią mówić po polsku?*]

  w [nazwa:onom Maćkowcach]

  w nocy pszychodźi tyn starosta do ńich każe: [npol: d'ity szo wy r`ob'ite …]

Lematyzacja

chudz’ema → chodzimy → **chodzić**

chudźema → chodzimy → **chodzić**

chodźema → chodzimy → **chodzić**

kub’ita → kobieta → **kobieta**

kub’ituf → kobiet → **kobieta**

Interpretacja zapisu

Przywrócenie konwencji polskiej ortografii

- b’a → bia
- ćo → cio

b’ały → biały
ćoća → ciocia
ćasto → ciasto

- ż → ż | rz

żeka → żeka | rzeka
żaba → żaba | rzaba

Interpretacja zapisu

Usunięcie zapisu specyficznych cech wymowy

- c’ → ć, ci

  c’ma → ćma

  c’oc’a → ciocia

- .e → e

  zboż.e → zboże

- y → y | e

  zboży → zboże | zboży

Interpretacja zapisu

Zastąpienie gwarowych końcówek gramatycznych literackimi

- -ma → -my chcema → chcemy
- -uf, -uw → - mamuf → mam

Interpretacja zapisu

Rezultaty zamian

kuńczyła ‘kończyła’ → kączeła, kączyła, kańczeła, kańczyła, kończeła, kończyła, kuńczeła, kuńczyła, kóńczeła, kóńczyła

m’es’c’e ‘mieście (miasto)’ → misci, miści, miźci, miscie, miście, miźcie, miesci, mieści, mieźci, miescie, mieście, mieźcie

gark’i ‘garnki’→ garki, garkie, garkij, garkiej

Wybór interpretacji

Porównanie rezultatów zamian z listą polskich form wyrazowych

- formy identyczne (kończyła; mieści, mieście)
- formy podobne (garki – garnki)

Wybór interpretacji

Preferowane wyniki bardziej podobne do postaci wyjściowej:

- **m’es’c’e** ‘mieście’ → misci, miści, miźci, miscie, miście, miźcie, miesci, mieści, mieźci, miescie, mieście, mieźcie

Wybór interpretacji – formy podobne

- garki → arki, barki, gacki, gadki, gaiki, ganki, garbi, garbki, garnki, gatki, gwarki, górki, jarki, karki, marki, ogarki, parki, tarki

- @append: 3
- @append: {splg} → 2
- @subst: e → ę → 1
- @subst: o → ą → 2
- @subst: {splg} → {sam} → 4
- @subst: {sam} → {splg} → 4

Wybór interpretacji – formy podobne

- garki → arki, barki, gacki, gadki, gaiki, ganki, garbi, garbki, garnki, gatki, gwarki, górki, jarki, karki, marki, ogarki, parki, tarki

- @append: 3
- @append: {splg} → 2
- @subst: e → ę → 1
- @subst: o → ą → 2
- @subst: {splg} → {sam} → 4
- @subst: {sam} → {splg} → 4

Lematyzacja

Morfeusz SIaT (Z. Saloni, M. Woliński)

- kuńczyła → kończyła → **kończyć**

- m’es’c’e → mieście → **miasto**
- m’es’c’e → mieście → **miesić**

- gark’i → garbki → **garbek**
- gark’i → garnki → **garnek**
- gark’i → gwarki → **gwarek**

Lematyzacja

porównanie z listą frekwencyjną (na podstawie korpusu SFPW, clip.ipipan.waw.pl/PL196x)

- kuńczyła → kończyła → **kończyć**

- m'es'c'e → mieście → **miasto**
- m'es'c'e → mieście → **miesić**

- gark'i → garbki → **garbek**
- gark'i → garnki → **garnek**
- gark'i → gwarki → **gwarek**

Uzupełnienie listy form wyrazowych

2 tys. leksemów dyferencjalnych

utworzenie hipotetycznych form tych leksemów:

- lista form hasłowych

  +

- tablica zakończeń form polskich (www.sjp.pl)

Uzupełnienie listy form wyrazowych

kycnąć/BeFIj
kinąć/BeFIj

doczka/mMN
rubaszka/mMN
pawozka/mMN

Maćkowce/fW

WZÓR (www.sjp.pl):
bryknąć/BeFIj

brzózka/mMN

skrzypce/fW

- *Maćkowce*, *Maćkowiec*, *Maćkowcom*…
  jak *skrzypce*, *skrzypiec*, *skrzypcom*…

Dalsze kroki

- dodanie kwalifikatora etymologicznego i znaczenia
- przystosowanie do wyszukiwarki Poliqarp

Zalety i wady programu

- dość skuteczny
- łatwe przygotowanie tekstów
- możliwość dostosowania do różnych podsystemów polszczyzny
- łatwe testowanie wyników na poszczególnych etapach

- trudny w obsłudze
- nie jest wspierany

Perspektywy

- korpus i słownik polszczyzny rejonu brasławskiego na Białorusi (kier. prof. J. Rieger)
- inne korpusy polszczyzny kresowej?