

# Automatyczna lematyzacja tekstu gwarowego w zapisie fonetycznym – program FonOrt

Aleksandra Wieczorek

Warszawa, 21 listopada 2013

# Przykład

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

# Wariantywność (fonetyczna i gramatyczna)

- chodzimy = chodźimy, chudźimy, ch.odźimy, chodz'imy, chodźema, chudźema...

# 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

- $b'a \rightarrow bia$                        $b'ały \rightarrow biały$
- $\dot{z} \rightarrow \dot{z} \mid rz$                        $\dot{z}eka \rightarrow \dot{z}eka \mid rzeka$   
                                                  $\dot{z}aba \rightarrow \dot{z}aba \mid rzaba$

# Interpretacja zapisu

Usunięcie zapisu specyficznych cech wymowy

- $c' \rightarrow \acute{c}$                        $c'ma \rightarrow \acute{c}ma$
- $.e \rightarrow e$                        $zbo\acute{z}.e \rightarrow zbo\acute{z}e$
- $y \rightarrow y \mid e$                        $zbo\acute{z}y \rightarrow \text{zbo\acute{z}e} \mid zbo\acute{z}y$

# 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 tych wyników, które istnieją w polszczyźnie ogólnej (porównanie z listą polskich form wyrazowych)

**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

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

# 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

# Lematyzacja

## Morfeusz SGJP

- kuńczyła → kończyła → **kończyć**
- m'es'c'e → mieście → **miasto**
- m'es'c'e → mieście → **miesiąc**
- 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](http://clip.ipipan.waw.pl/PL196x))

- kuńczyła → kończyła → **kończyć**
- m'es'c'e → mieście → **miasto**
- m'es'c'e → mieście → **miesiąc**
- gark'i → garbki → **garbek**
- gark'i → garnki → **garnek**
- gark'i → gwarki → **gwarek**

# Wyrazy nieogólnopolskie

utworzenie hipotetycznych form tych leksemów (np. *doczka* – *doczki*, *doczce*, *doczkę*, *doczką*):

- lista form hasłowych
- +
- tablica zakończeń form polskich ([www.sjp.pl](http://www.sjp.pl))

# Uzupełnienie listy form wyrazowych

WZÓR ([www.sjp.pl](http://www.sjp.pl)):

doczka/mMN

rubaszka/mMN

pawozka/mMN

brzózka/mMN

kycnać/BeFIj

kinąć/BeFIj

bryknąć/BeFIj

Maćkowce/fW

skrzypce/fW

- *doczki, doczce, doczkę, doczek...*  
*jak brzózki, brzózce, brzózkę, brzózek...*

# 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