PAN IJP
Institute of Polish Language
Polish Academy of Sciences

Tagset adaptation to language changing over time. The case of the Electronic Corpus of the 17th- and 18th-century Polish Texts

Aleksandra Wieczorek

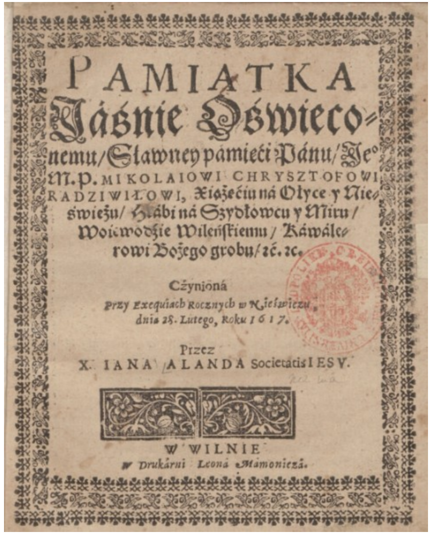
THE ELECTRONIC CORPUS OF 17TH- AND 18TH-CENTURY POLISH TEXTS

- Cryptonym: KORBA (KORpus BARokowy ‘baroque corpus’)
- Principal investigator: Włodzimierz Gruszczyński
- Team: members of Institute of Polish Language and Institute of Computer Science, Polish Academy of Sciences

- Funding: Polish Ministry of Science and Higher Education, National Programme for the Development of Humanities grant (project number 0036/NPRH2/H11/81/2012 and 0413/NPRH7/H11/86/2018)
- Project duration: 2013-2018 and 2019-2023
- Coordinating body: Institute of Polish Language, Polish Academy of Sciences
- Cooperation: Institute of Computer Science, Polish Academy of Sciences

www.korba.edu.pl

THE ELECTRONIC CORPUS OF 17TH- AND 18TH-CENTURY POLISH TEXTS



PAMIATKA
Jáśnie Oświeco-
nemu/ Sławney pámięci Pánu/ Je°
M. p. MIKOLAIOWI CHRYSZTOFOWI
RADZIWIŁOWI, Xiążęćiu ná Olyce y Nie-
świeżu/ Hrabi ná Szydłowcu y Miru/
Woiewodźie Wileńſkiemu/ Káwále-
rowi Bożego grobu/ ꝛc. ꝛc.

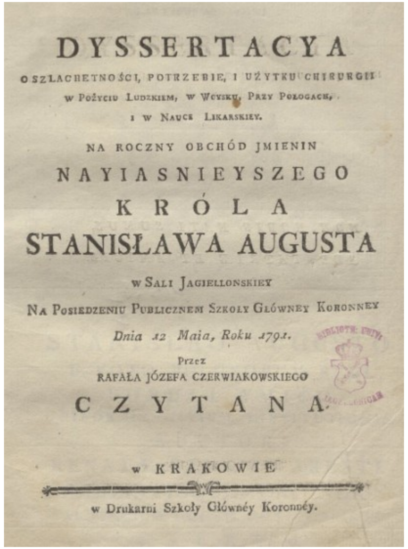
Czyniona
Przy Exequiach Rocznych w Nieświeżu,
dnia 28. Lutego, Roku 1617.

Przez
X. IANA ALANDA Societatis IESV.

W WILNIE
w Drukárni Leoná Mámonicżá.

1617

- 13.5M → 25M tokens
- richly annotated
- usage: research of grammar and lexis of 17th—18th centuries
- 1601—1800 (Baroque, Enlightenment)



DYSSERTACYA
O SZLACHETNOŚCI, POTRZEBIE, I UŻYTKU CHIRURGII
w Pożyciu Ludzkiem, w Woysku, Przy Połogach,
i w Nauce Likarskiey.

NA ROCZNY OBCHÓD JMIENIN
NAYIASNIEYSZEGO
KRÓLA
STANISŁAWA AUGUSTA
w Sali Jagiellonskiey
Na Posiedzeniu Publicznem Szkoły Główney Koronney
Dnia 12 Maia, Roku 1791.
Przez
RAFAŁA JÓZEFA CZERWIAKOWSKIEGO
CZYTANA.

w KRAKOWIE
w Drukarni Szkoły Główney Koronney.

1791

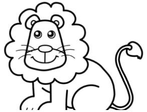
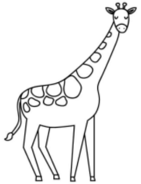
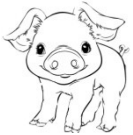
PAN IJP Institute of Polish Language
Polish Academy of Sciences

Changes in Polish grammar during 17th and 18th c.

- Appearance of new grammatical categories or their values (e.g. of "masculine-personality" category)
- Disappearance of some grammatical categories or their values (e.g. disappearance of dual number)
- Changes of inflectional paradigms (e.g. of numerals)
- Changes of inflectional endings
- …

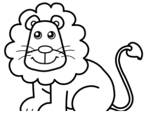
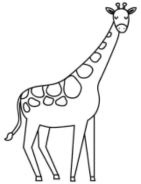
1. The new category: masculine-personality

Grammatical cathegory of gender in Polish – nouns

m (masculine)		**f** (feminine)		**n** (neuter)	
pan ‘gentleman’		królowa ‘queen’		dziecko ‘child’	
lew ‘lion’		żyrafa ‘giraffe’		prosię ‘piglet’	
dom ‘house’		harfa ‘harp’		okno ‘window’	

PAN IJP Institute of Polish Language Polish Academy of Sciences

Grammatical cathegory of gender in Polish – adjectives

m	f	n
dobry ‘good’:	dobra:	dobre:
pan	królowa	dziecko
lew 	żyrafa 	prosię
dom	harfa	okno

PAN IJP Institute of Polish Language Polish Academy of Sciences

3 masculine subgenders in modern Polish

- personal = m1

- animate impersonal = m2

- inanimate = m3

PAN IJP Institute of Polish Language Polish Academy of Sciences

3 masculine subgenders in modern Polish

sg. N pan m1

…

A pan-a m1

…

pl. N pan-owie m1

…

A pan-ów m1

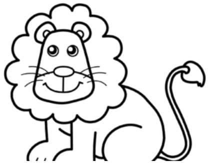
…

sg. N lew m2

…

A lw-a m2

…



pl. N lw-y m2

…

A lw-y m2

…

sg. N dom m3

…

A dom m3

…

pl. N dom-y m3

…

A dom-y m3

…

PAN IJP Institute of Polish Language Polish Academy of Sciences

3 masculine subgenders in modern Polish

sg. N pan m1

…

A pan-a m1

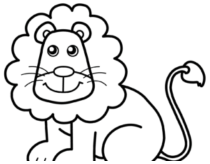
…

sg. N lew m2

…

A lw-a m2

…



sg. N dom m3

…

A dom m3

…

pl. N pan-owie m1

…

A pan-ów m1

…

pl. N lw-y m2

…

A lw-y m2

…

pl. N dom-y m3

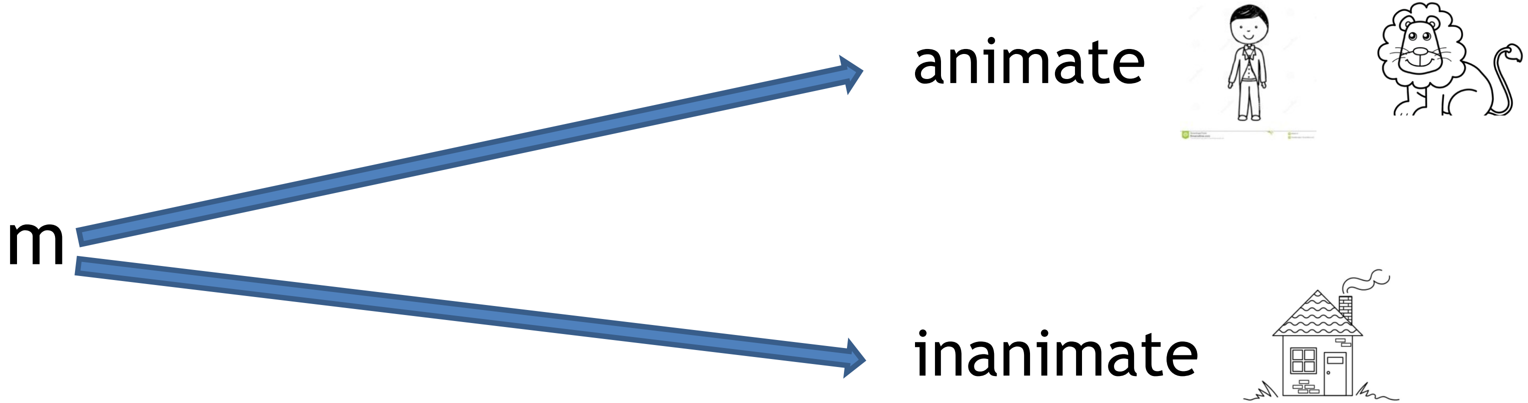
…

A dom-y m3

…

PAN IJP Institute of Polish Language Polish Academy of Sciences

Changes of the inflection of masculine nouns in Old Polish (up to 15th c.)



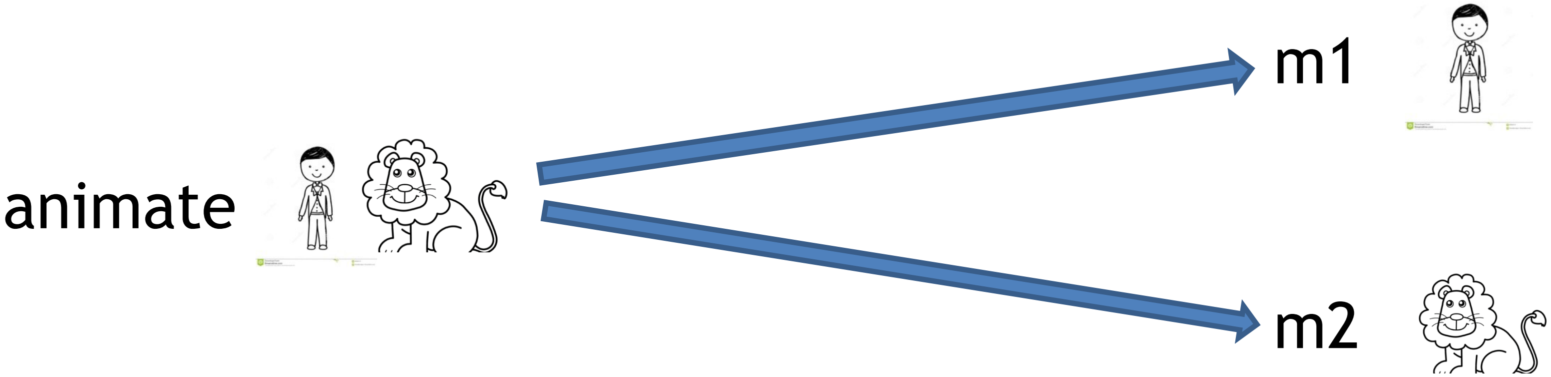
m

animate

inanimate

PAN IJP Institute of Polish Language Polish Academy of Sciences

Changes of the inflection of masculine nouns in 17th—18th centuries



animate

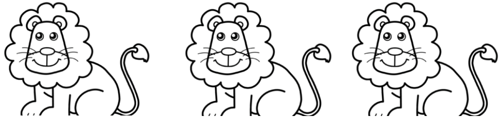
m1

m2

PAN IJP Institute of Polish Language Polish Academy of Sciences

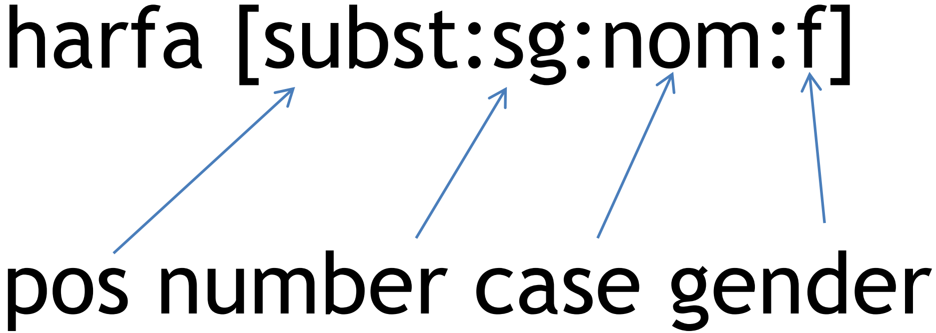
Changes of the inflection of masculine nouns in 17th—18th centuries

- N. pl pan-y / pan-owie
- N. pl lw-y / lw-owie



Tagset

A set of markers signifying part of speech and morpho-syntactic features specific for this part of speech.



harfa [subst:sg:nom:f]

pos number case gender

Tagging the gender in modern Polish

The gender is assigned to the particular noun, e.g.

pan [subst:sg:nom:m1]

pan-owie ‘gentlemen’
[subst:pl:nom:m1]

lew [subst:sg:nom:m2]

lw-y ‘lions’
[subst:pl:nom:m2]

How to tag the gender in Middle Polish?

pan [subst:sg:nom:m1]

pan-owie / pan-y
[subst:pl:nom:m?]

lew [subst:sg:nom:m2]

lw-owie / lw-y
[subst:pl:nom:m?]

lwowie

PAN IJP Institute of Polish Language Polish Academy of Sciences

Tagging the gender in KorBa

pan [subst:sg:nom:m]

pan-a [subst:sg:gen:m]

pan-owie
    [subst:pl:nom:manim1]

pan-y [subst:pl:nom:m]

lew [subst:sg:nom:m]

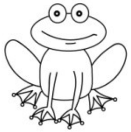
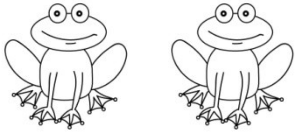
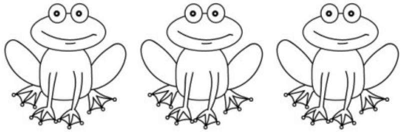
lw-a [subst:sg:gen:m]

lw-owie
    [subst:pl:nom:manim1]

lw-y [subst:pl:nom:m]

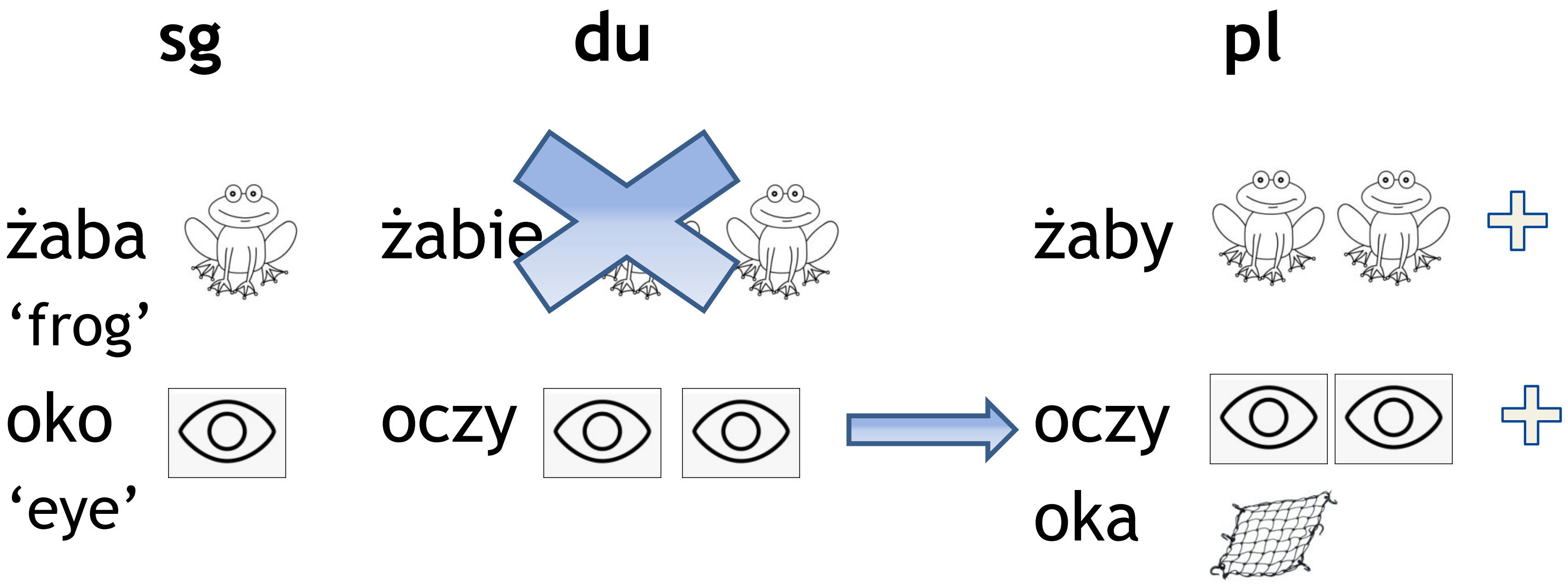
2. The disappearance of dual number

Category of number in Old Polish

sg	du	pl
żaba  ‘frog’	żabie 	żaby 
oko 👁 ‘eye’	oczy 👁 👁	oka 👁 👁 👁

PAN IJP Institute of Polish Language Polish Academy of Sciences

The disappearance of dual during 17th—18th c.



sg	du	pl
żaba ‘frog’	żabie	żaby
oko ‘eye’	oczy	oczy oka

How to annotate forms like “oczy” in KorBa?

Up to 1740

oczy → du

oka → pl

After 1740

oczy → pl

oka → pl

Conclusions

- The development of the grammatical system should be reflected in the morphosyntactic annotation of the corpus.
- There is no universal solution – for each grammatical feature, different issues should be taken into account:
  - substantive issues, e.g. the degree of prevalence of a given linguistic phenomenon; the timeframe for linguistic change
  - practical issues, e.g. the possibility of creating useful corpus queries and conducting linguistic research in future.

Thank you!

www.korba.edu.pl