Enhanced digital editions: retrieving POS tags from pre-digital word indexes

Joanna Byszuk, Magdalena Król, Maciej Eder

10 Computational 01
01 Stylistics 0101000
11 Group 011010110

EADH 2018 Galway

PAN IJP Institute of Polish Language Polish Academy of Sciences

Why?

Biblioteka Pisarzów Polskich – A Collection of Polish Writers: context

About Biblioteka Pisarzów Polskich

Two series: A and B:

34 volumes in B series, covering literary heritage of most important Old and Middle Polish writers

Linguistic edition by best scholars, printed for over 70 years, now forgotten or really unique (thus, highly priced)

Consequent framework supported by years of textual analysis

About Biblioteka Pisarzów Polskich – construction

- images of the original sources
- a philological transcription and a commentary
- comprehensive indexes of all the words used in the edited texts:
  - their grammatical tags,
  - the information about inflection (if applicable),
  - precise location of the form in question in the original text

– > complete sets of POS tags identified manually by cutting-edge scholars in Polish philology

Posłow Graeckich.

Jesli namniey przewinył/ być mu w okowách.
Więc ia podobno z mnieyszym niebespieczeństwem
Grzeszę: bo sam sie trácę swym wszeteczeństwem.
Przełożonych występy miástá zgubiły/
Y szerokie do gruntu Cárstwá zniszczyły.

POSEŁ. HELENA.

PO. Dobrą nowinę pániey swoiey niosę/
Rozumiem temu/ że iuż dawno tego
Poselstwá czeka/ serce swe troskámi
Y płáczem trapiąc: Ale oto práwie
Ná czás wychodzi z domu. O Krolowa/
Wdzięczney nowiny posłá masz przed sobą.
Hel. Day Boże/ byś co przyniosł pocieszngo.
Po. Posłowie twoi iáko przyiechali/
Ták odieżdżáią/ á ty przedsie známi.
Hel. Byłeś sam w rádzie/ czyś słyszał od kogo?
Po. Byłem przy wszytkim: y prosto mi ztámtąd
Iść ALEXánder do ciebie roskazał.
Hel. Jeszczeć niewidzę/ szczegobych sie práwie
Ucieszyć miáłá/ wszákże powiedz przedsie
Jáko co było. Po. Powiem/ ieno słuchay.
Skoro w rádzie zásiedli pánowie/ Krol naprzod
Tę rzecz do nich vczynił. Niezwykłem nic nigdy
Bez rády wászey czynić: ábych też zwykł kiedy/
(czeg w pámięci niemam) w tey spráwie koniecznie
Synà swego bych niechciał: áby mie oycowska

180 Jesli namniej przewinił – być mu w okowách.
Więc ja podobno z mniejszym niebezpieczeństwem Grzeszę, bo sam sie trácę swym wszeteczeństwem. Przełożonych występy miástá zgubiły
I szerokie do gruntu cárstwá zniszczyły. 5

POSEŁ – HELENA

PO.: Dobrą nowinę pániej swojej niosę.
Rozumiem temu, że już dawno tego Poselstwá czeka, serce swe troskámi
I płáczem trapiąc. Ale oto práwie 10
185 Ná czás wychodzi z domu. O, krolowa, Wdzięcznej nowiny posłá masz przed sobą.
HEL.: Daj boże, byś co przyniosł pocieszngo.
PO.: Posłowie twoi jáko przyjechali,
Ták odjeżdżają, á ty przedsie z námi! 15
190 HEL.: Byłeś sam w rádzie, czyś słyszał od kogo?
PO.: Byłem przy wszytkim i prosto mi stámtąd Iść Aleksánder do ciebie rozkazał.
HEL.: Jeszczeć nie widzę, *z* czego bych sie práwie Ucieszyć miáłá, wszákże powiedz przedsię 20
195 Jáko co było. PO.: Powiem, jeno słuchaj.
Skoro w rádzie zásiedli pánowie, krol naprzod
Tę rzecz do nich uczynił: „Nie zwykłem nic nigdy Bez rády wászej czynić; á bych też zwykł kiedy
(Czego w pámięci nie mam), w tej spráwie koniecznie 25
200 Synà swego bych nie chciał, áby mie ojcowska
B3 Miłość

Why is grammar tagging of old texts hard?

Difficult disambiguation of word sense:

Big number of word novelties:

- atypical word formation,
- improper inflection adjusted to rhyme and melody of a poem,
- fluency with morphology – free switching between categories

Why is grammar tagging of old texts hard?

Non-trivial grammatical problems:

- anaphors,
- elipsis,
- non-neutral word order (used for different purposes)

Non-obvious cultural contexts in proper names, unknown properties of common nouns, aspects of verbs etc.

Contemporary applications of linguistic resources from old editions

Contemporary applications of linguistic resources (not only) from old editions

Unflattening existing critical editions into digital ones

Diachronic investigations

Contemporary applications of linguistic resources (not only) from old editions

- Possible training set for automatic tagging of historical texts (yields a large set of manually annotated data)
- More precise description of grammatical trends (instead of descriptions of linguistic phenomena) – quantitative input needed in historical data
- An enhancement for Universal Dependencies

Why Kochanowski to start with?

Jan Kochanowski

A writer with good ear for dialects and spoken language, using a sublime and sophisticated Polish at the same time:

- mixed registers,
- interesting word formation,
- playing with the sound and the sense of words

Wide range of topic covered in his works.

Jan Kochanowski’s heritage

A strong influence on the development of Polish literature.

Visible presence in modern culture and education.

Multiple resources are now prepared, including:

- dictionaries,
- critical editions,
- indexes,
- theoretical descriptions.

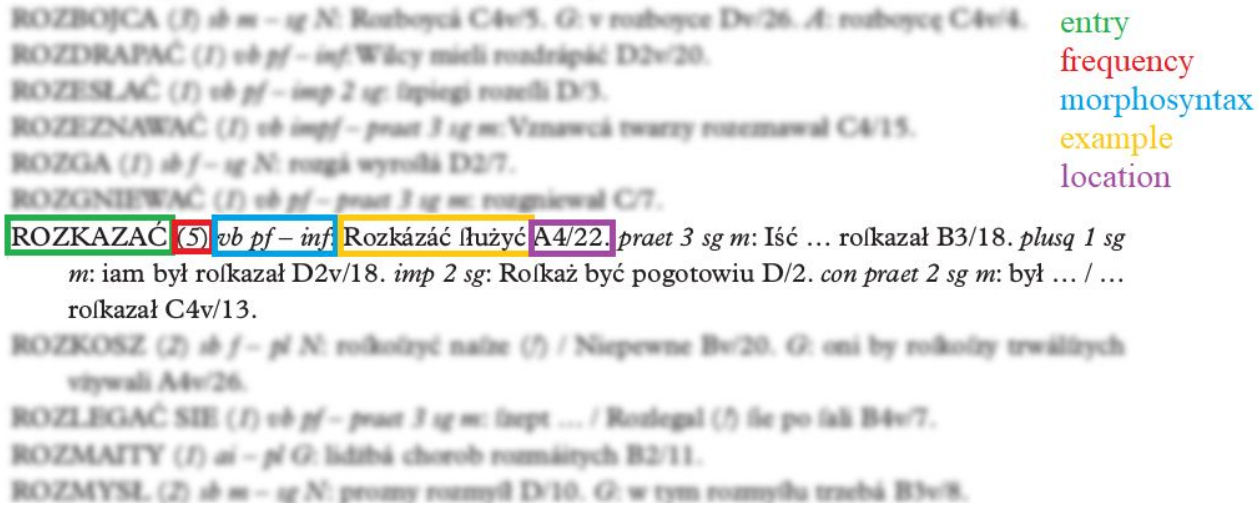
Odprawa posłów greckich

- Manuscript created around 1577,
- Many editions preserved, until 1877 printed in collected works (at least 15 different versions)
- Written for the wedding celebrations of Jan Zamoyski and Krystyna Radziwiłłówna
- A typical example of Kochanowski's works: Mix of archaic and innovative vocabulary and structures

Unflattening print critical editions

Unflattening print critical editions – challenges

- Editions printed between 1889 (1952) and now
  - obtaining them
- Cracking the case(s) open
  - From a printed index to a usable structure

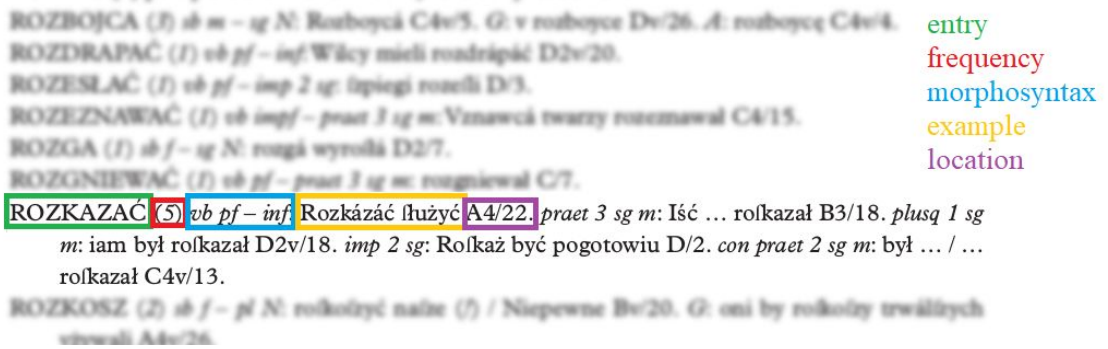


Challenge: getting the data

- Availability / Copyright

OCR problems:

- Various print
- Various state of preservation
- Typesetting – e.g. italics matter, delimitation hard to catch automatically



Unflattening print critical editions – steps

1. OCR and cleaning
2. Identifying elements in each entry
3. Extracting elements
4. Verifying index of elements
5. Building tags
6. Tagging text
7. Verifying accuracy & comparing with automatic taggers

Conclusions

Conclusions

Possibilities:

- Easier verification of text tagging.
- Tagging independent of transliteration and transcription.
- With more ready editions we can unify grammar description within the whole corpus.

Thank you!