Morphosyntactic Annotation of Historical Texts

The Making of the Baroque Corpus of Polish

W. Kieraś, D. Komosińska, E. Modrzejewski & M. Woliński

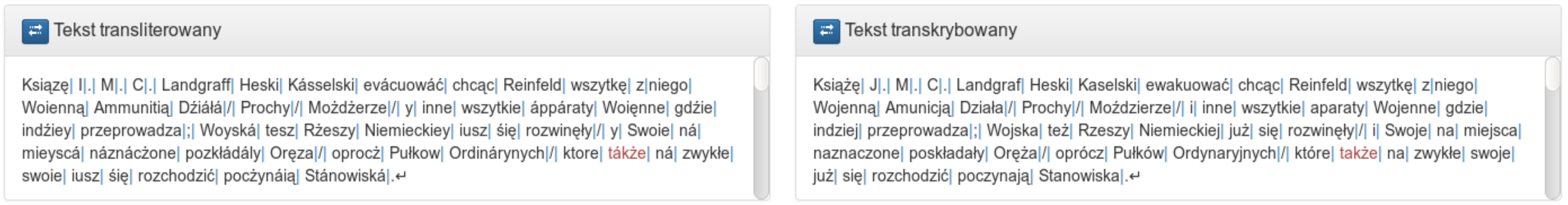
Institute of Computer Science & Institute of Polish Language
Polish Academy of Sciences

Abstract

We present some technical issues concerning processing 17th & 18th century texts for the purpose of building a corpus of that period. We describe a chain of procedures leading from transliterated source texts to morphological annotation of text samples that was implemented for building the Baroque Corpus of Polish. The procedure consists of: automatic transliteration from original spelling to modern one, morphological analysis (including the construction of an inflectional dataset for Baroque Polish) and a tool for manual morphosyntactic annotation. The toolchain is being used to create a small manually validated subcorpus, which will serve as training data for a stochastic tagger. Then a larger corpus will be annotated automatically and made available via the Poliqarp corpus search tool.

Automatic transcription

- The source texts of the Baroque Corpus of Polish divide into three types of editions: (1) original, (2) 19th century, and (3) contemporary.
- Types (1), (2) require automatic transcription to normalize the texts to make them better suited for morphological analysis.
- The converter contains nearly 4.000 substitution rules, based on regular expressions and the context of the sequence of characters' appearance.
- To check and correct the transcription, corpus annotators have a possibility to work on both transliterated and transcribed versions:



Tekst transliterowany

Książę| I|.| M|.| C|.| Landgraff| Heski| Kásselski| evácuowáć| chcąc| Reinfeld| wszytkę| z|niego| Woienną| Ammunitią| Dźiáłá|/| Prochy|/| Możdżerze|/| y| inne| wszytkie| ápparáty| Woięnne| gdźie| indźiey| przeprowadza|;| Woyská| tesz| Rżeszy| Niemieckiey| iusz| śię| rozwinęły|/| y| Swoie| ná| mieyscá| náznáczone| pozkłádály| Oręza|/| oprocż| Pułkow| Ordinárynych|/| ktore| tákże| ná| zwykłe| swoie| iusz| śię| rozchodzić| poczynáią| Stánowiská|.↵

Tekst transkrybowany

Książę| J|.| M|.| C|.| Landgraf| Heski| Kaselski| ewakuować| chcąc| Reinfeld| wszytkę| z|niego| Wojenną| Amunicją| Działa|/| Prochy|/| Moździerze|/| i| inne| wszytkie| aparaty| Wojenne| gdzie| indziej| przeprowadza|;| Wojska| też| Rzeszy| Niemieckiej| już| się| rozwinęły|/| i| Swoje| na| miejsca| naznaczone| poskładały| Oręża|/| oprócz| Pułków| Ordynaryjnych|/| które| także| na| zwykłe| swoje| już| się| rozchodzić| poczynają| Stanowiska|.↵

A text sample manually transliterated (left) and then automatically transcribed (right)

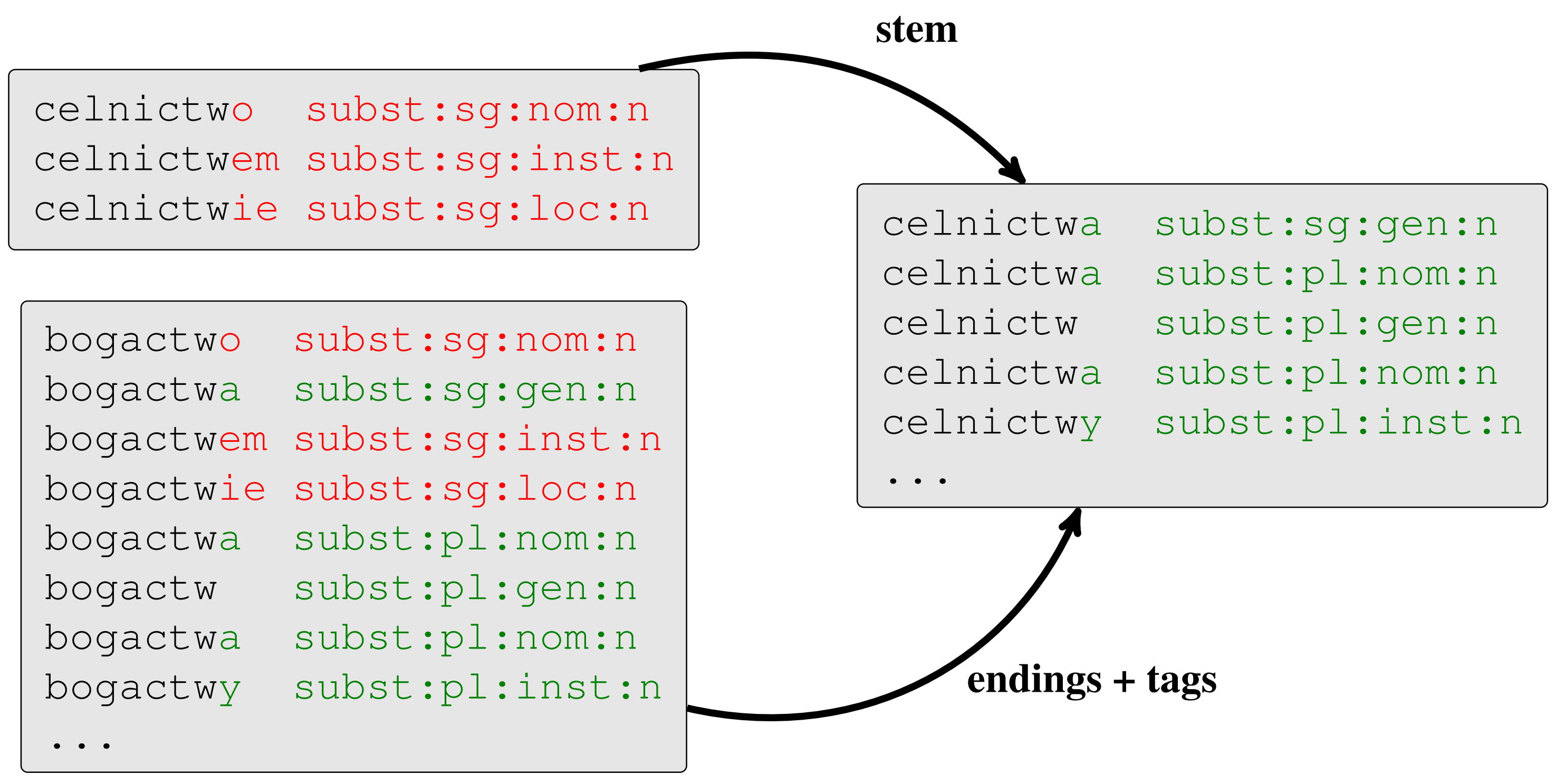
- Automatic transcription normalizes the tokens, greatly reducing their types, and causes the number of tokens unknown to the automatic morphological analyzer (see below) to plummet to 5.4%:

	transliteration	transcription
token occurrences	12,814,830	12,832,214
token types	646,410	476,733
unrecognized token types	75.55%	48.08%
unrecognized token occurrences	24.04%	5.4%

Number of tokens in the two representations of the corpus

Morphological Analysis

- Automatic morphological analysis is performed using Morfeusz analyzer with a dictionary based on the following sources of data:
  - inflectional information from the Electronic Dictionary of 17th & 18th century Polish (e-SXVII, `http://sxvii.pl/`)
  - contemporary data of Grammatical Dictionary of Polish (SGJP) modified ("aged") to fit into the tagset for Baroque morphosyntax.
- Some automatically obtained extensions of inflectional paradigms from e-SXVII was also used. The extension procedure enriched the data set by over 210,000 new forms.



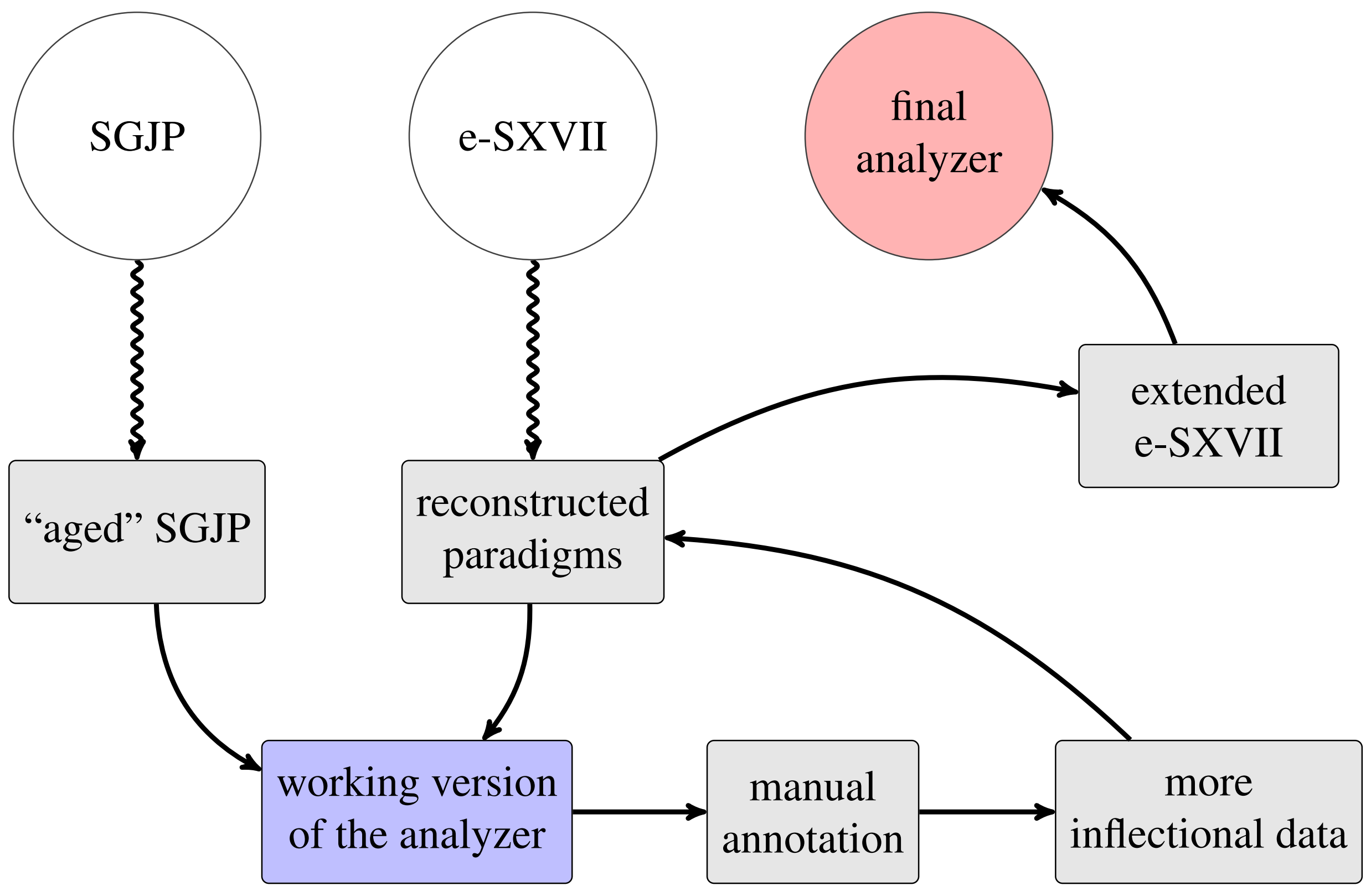
```
celnictwo  subst:sg:nom:n
celnictwem subst:sg:inst:n
celnictwie subst:sg:loc:n
```

```
bogactwo  subst:sg:nom:n
bogactwa  subst:sg:gen:n
bogactwem subst:sg:inst:n
bogactwie subst:sg:loc:n
bogactwa  subst:pl:nom:n
bogactw   subst:pl:gen:n
bogactwa  subst:pl:nom:n
bogactwy  subst:pl:inst:n
...
```

```
celnictwa  subst:sg:gen:n
celnictwa  subst:pl:nom:n
celnictw   subst:pl:gen:n
celnictwa  subst:pl:nom:n
celnictwy  subst:pl:inst:n
...
```

Example of partial reconstruction of inflectional paradigm from e-SXVII

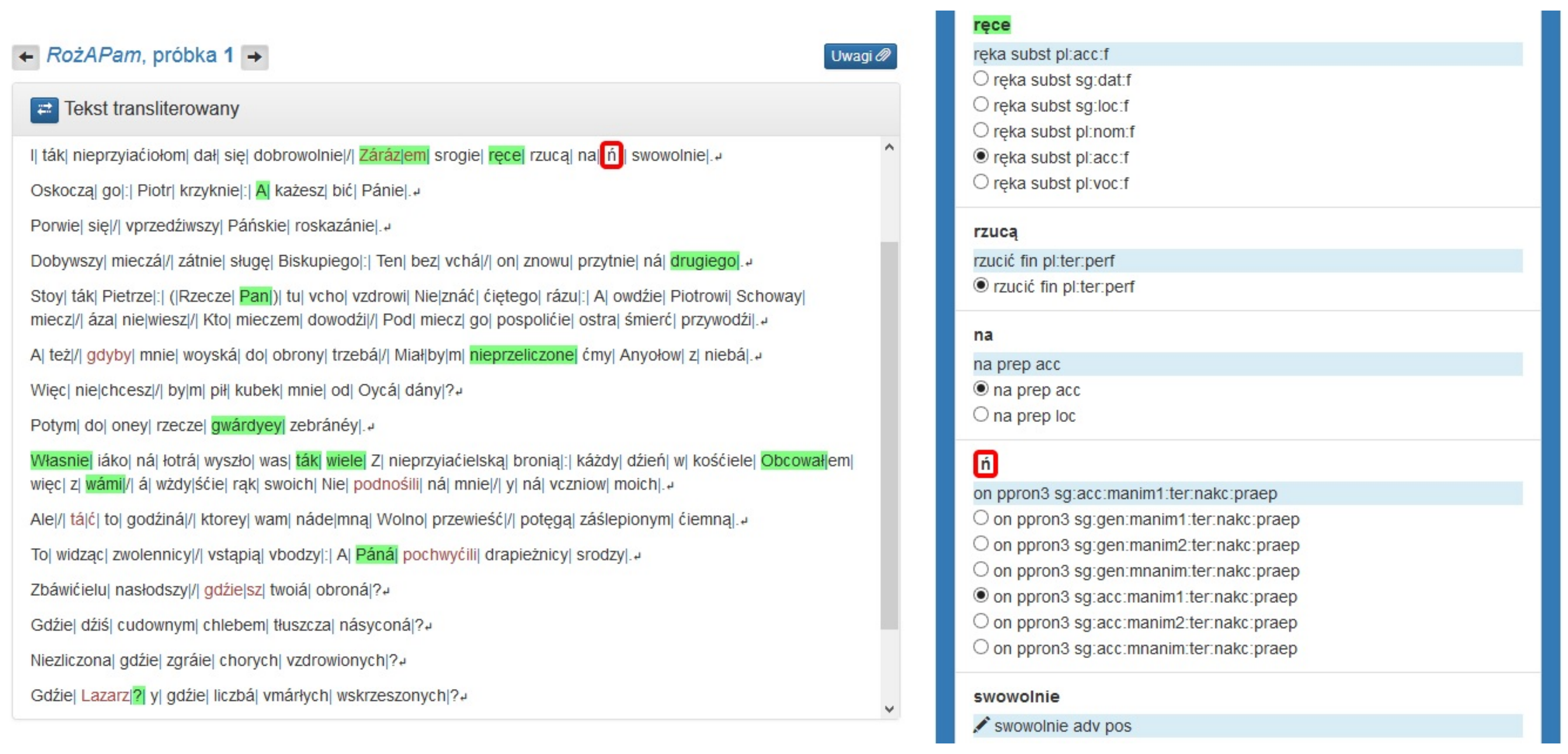
- "Aging" SGJP and reconstructing forms unnoted in e-SXVII allowed to reduce the number of tokens unrecognized by the analyzer in the corpus from 11% to 5,4%.
- Tokenization rules also needed to be adjusted. Baroque Polish, when compared to modern language, is characterized by wider use of joint spelling. Numerous clitics and particles are much more mobile and can be attached to a wider range of forms than nowadays.
- The analyzer is in the constant development based on the anotators' feedback. Its final version will be used for tagging the full corpus.



Workflow in the process of building the morphological analyzer

Manual Annotation

- A corpus of 500,000 tokens of BCP will be manually annotated (currently ca. 150,000).
- All tokens marked as punctuation, foreign elements or structural markers were excluded for the purpose of the evaluation.
- The percentages of manual changes introduced by annotators are as follows:
  - modification of transcription – 2.66% of tokens, of tokenization – 1,59%,
  - morphological interpretation modified by the annotator – 8.6%.
- Each text sample is annotated independently by two annotators. In case of conflict the adjudicator has to resolve it by selecting one of the existing annotations or providing a new one.
- Currently annotators generate conflicts on 12.6% of tokens.



Tekst transliterowany

I| ták| nieprzyiaćiołom| dał| się| dobrowolnie|/| Záráz|em| srogie| ręce| rzucą| na| ń| swowolnie|.↵

Oskoczą| go|:| Piotr| krzyknie|:| A| każesz| bić| Pánie|.↵

Porwie| się|/| vprzedźiwszy| Páńskie| roskazánie|.↵

Dobywszy| mieczá|/| zátnie| sługę| Biskupiego|:| Ten| bez| vchá|/| on| znowu| przytnie| ná| drugiego|.↵

Stoy| ták| Pietrze|:| (|Rzecze| Pan|)| tu| vcho| vzdrowi| Nie|znáć| ćiętego| rázu|:| A| owdźie| Piotrowi| Schoway| miecz|/| áza| nie|wiesz|/| Kto| mieczem| dowodźi|/| Pod| miecz| go| pospolićie| ostra| śmierć| przywodźi|.↵

A| też|/| gdyby| mnie| woyská| do| obrony| trzebá|/| Miał|by|m| nieprzeliczone| ćmy| Anyołow| z| niebá|.↵

Więc| nie|chcesz|/| by|m| pił| kubek| mnie| od| Oycá| dány|?↵

Potym| do| oney| rzecze| gwárdyey| zebránéy|.↵

Własnie| iáko| ná| łotrá| wyszło| was| ták| wiele| Z| nieprzyiaćielską| bronią|:| káżdy| dźień| w| kośćiele| Obcował|em| więc| z| wámi|/| á| wżdy|śćie| rąk| swoich| Nie| podnośili| ná| mnie|/| y| ná| vczniow| moich|.↵

Ale|/| tá|ć| to| godźiná|/| ktorey| wam| náde|mną| Wolno| przewieść|/| potęgą| záślepionym| ćiemną|.↵

To| widząc| zwolennicy|/| vstąpią| vbodzy|:| A| Páná| pochwyćili| drapieżnicy| srodzy|.↵

Zbáwićielu| nasłodszy|/| gdźie|sz| twoiá| obroná|?↵

Gdźie| dźiś| cudownym| chlebem| tłuszcza| násyconá|?↵

Niezliczona| gdźie| zgráie| chorych| vzdrowionych|?↵

Gdźie| Lazarz|?| y| gdźie| liczbá| vmárłych| wskrzeszonych|?↵

ręce
ręka subst pl:acc:f
○ ręka subst sg:dat:f
○ ręka subst sg:loc:f
○ ręka subst pl:nom:f
● ręka subst pl:acc:f
○ ręka subst pl:voc:f

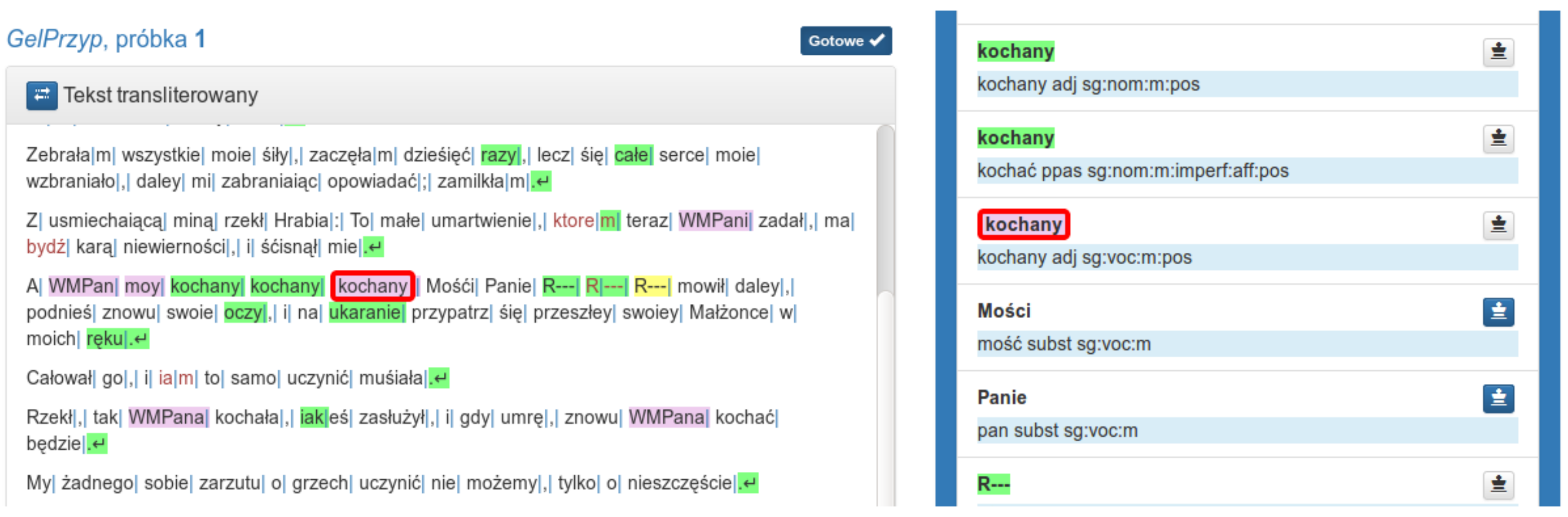
rzucą
rzucić fin pl:ter:perf
● rzucić fin pl:ter:perf

na
na prep acc
● na prep acc
○ na prep loc

ń
on ppron3 sg:acc:manim1:ter:nakc:praep
○ on ppron3 sg:gen:manim1:ter:nakc:praep
○ on ppron3 sg:gen:manim2:ter:nakc:praep
○ on ppron3 sg:gen:mnanim:ter:nakc:praep
● on ppron3 sg:acc:manim1:ter:nakc:praep
○ on ppron3 sg:acc:manim2:ter:nakc:praep
○ on ppron3 sg:acc:mnanim:ter:nakc:praep

swowolnie
swowolnie adv pos

A sample being annotated as seen by an annotator in Anotatornia 2



Tekst transliterowany

Zebrała|m| wszystkie| moie| śiły|,| zaczęła|m| dziesięć| razy|,| lecz| śię| całe| serce| moie| wzbraniało|,| daley| mi| zabraniaiąc| opowiadać|;| zamilkła|m|.↵

Z| usmiechaiącą| miną| rzekł| Hrabia|:| To| małe| umartwienie|,| ktore|m| teraz| WMPani| zadał|,| ma| bydź| karą| niewierności|,| i| ścisnął| mie|.↵

A| WMPan| moy| kochany| kochany| kochany| Mośći| Panie| R---| R|---| R---| mowił| daley|,| podnieś| znowu| swoie| oczy|,| i| na| ukaranie| przypatrz| śię| przeszłey| swoiey| Małżonce| w| moich| ręku|.↵

Całował| go|,| i| ia|m| to| samo| uczynić| muśiała|.↵

Rzekł|,| tak| WMPana| kochała|,| iak|eś| zasłużył|,| i| gdy| umrę|,| znowu| WMPana| kochać| będzie|.↵

My| żadnego| sobie| zarzutu| o| grzech| uczynić| nie| możemy|,| tylko| o| nieszczęście|.↵

kochany
kochany adj sg:nom:m:pos

kochany
kochać ppas sg:nom:m:imperf:aff:pos

kochany
kochany adj sg:voc:m:pos

Mości
mość subst sg:voc:m

Panie
pan subst sg:voc:m

R---

Conflicts to be resolved as seen by an adjudicator

Conclusions & further work

- The BCP is a work in progress since the manual annotation of the subcorpus is ongoing.
- All components involved in processing text samples are being constantly enhanced according to the feedback from annotators and adjudicators, but it seems that the tools are already quite stable.
- When the manually annotated corpus is ready, a stochastic tagger will be trained on the annotations and the rest of the BCP corpus will be automatically annotated. We also plan to check whether machine learning techniques can be used to train a better transcriber based on transcriptions manually corrected by the annotators.
- The presented toolchain will also be used to prepare a similarly organized corpus of 19th century Polish texts.

Acknowledgements

The work being reported was co-financed by a National Science Centre, Poland grant DEC-2014/15/B/HS2/03119 and a Ministry of Science and Higher Education National Programme for the Development of Humanities grant 0036/NPRH2/H11/81/2012.

20th International Conference TEXT, SPEECH AND DIALOGUE, 27th–31st August 2017, Prague, Czechia