

Use of data from the electronic historical dictionary to build a corpus

Renata Bronikowska^{*}
Włodzimierz Gruszczyński^{*}
Maciej Ogrodniczuk^{**}
Marcin Woliński^{***}

^{*}Institute of Polish Language, Polish Academy of Sciences

^{**}Institute of Computer Science, Polish Academy of Sciences

Presentation prepared within the project
The electronic corpus of the 17th and 18th century Polish texts (up to 1772)
funded by National Programme for the Development of Humanities (NPRH)
for years 2013-2018

e-SXVII

i.e. *The electronic dictionary of Polish language of the 17th and 18th century*

- **Contractor:** Institute of Polish Language, Polish Academy of Sciences
- Since 2004 available on a regular basis at:
<http://sxvii.pl>
- Planned size: ca 80K entries
- Now contains ca 20K lexemes (of different status)
- Ca 52K forms – paradigms with blanks (only forms found in examined sources)
- Transliterated quotations from ca 850 sources

KORBA (KORpus BArokowy)

i.e. *The electronic corpus of the 17th and 18th century Polish texts (up to 1772)*

- Duration of the project: 2013-2018
- **Contractors:** Institute of Polish Language and Institute of Computer Science, Polish Academy of Science
- Planned size: 12M tokens
- Manual morphosyntactic annotation: 0.5M tokens
- Rich structural annotation (e.g. quotations assigned to the appropriate page in the text)
- Annotation of foreign language fragments (different tags assigned to different languages)
- Texts transliterated according to the rules accepted in e-SXVII

Corpus for dictionary

- Corpus allows to find quickly a word with its context. This helps to:
 - find orthographic, phonetic & morphological variants of the lexeme
 - establish its syntactic and lexical collocations
 - determine its meaning or meanings.
- Information about the texts contained in the corpus **enable(s)** to:
 - check frequency, chronology, geography etc. of the lexeme
 - determine **genres** of texts in which the lexeme is usually used
 - point out authors, who often use the lexeme.
- Quotations from the corpus serve as an illustration of the word meanings, idioms or collocations.

Dictionary for corpus

- Morphosyntactic tagging enhances possibilities of corpus searching
- Tools which support morphosyntactic tagging
 - inflectional analyser – ascribes all possible interpretations to analysed grammatical form
 - tagger – disambiguates grammatical interpretation
- Inflectional analyser is based on the grammatical dictionary, which contains lexemes with their paradigms
- Polish contemporary lexemes' paradigms are gathered in the *Grammatical Dictionary of Polish* (SGJP – <http://sgjp.pl>)
- SGJP constitutes a base for inflectional analyser Morfeusz, which supported tagging of National Corpus of Polish (NKJP – <http://nkjp.pl>)

Historical dictionary for historical corpus

BANKA *rzecz. f*

Warianty fonetyczne: BAŃKA, BANKA

Słowniki notują

Formy gramatyczne:

	<i>lp</i>
<i>M.</i>	banka bańka
<i>D.</i>	bańki
<i>C.</i>	bańce
<i>B.</i>	bańkę bańkę
<i>N.</i>	banką bańką
<i>Ms.</i>	bańce

	<i>lm</i>
<i>M.</i>	banki bańki
<i>D.</i>	baniek
	banków
<i>B.</i>	banki bańki
<i>N.</i>	bańkami
<i>Ms.</i>	bańkach

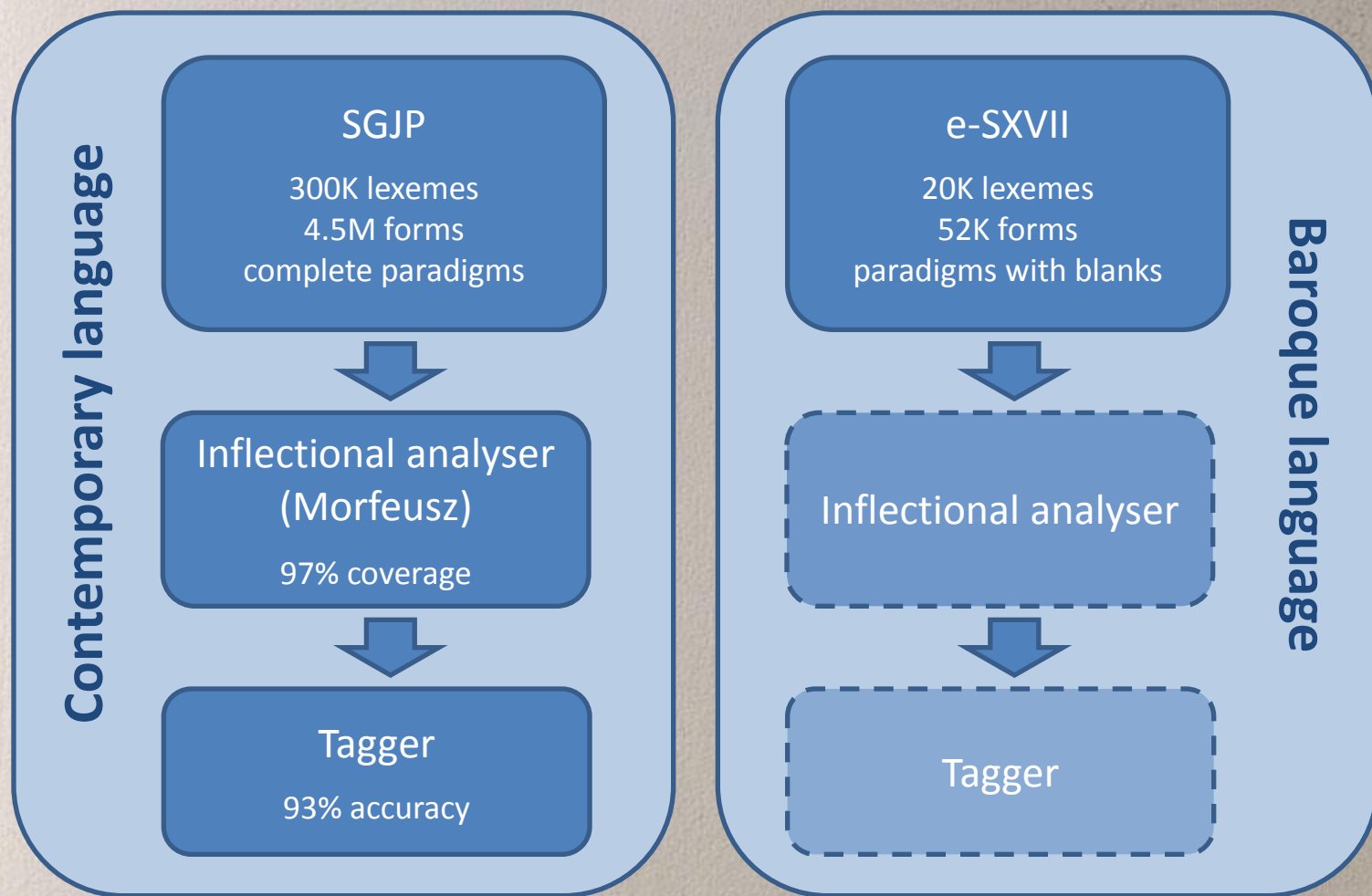
	<i>lpdw</i>
<i>M.</i>	bańce

1. »pękate naczynie ze szkła, drewna lub innego materiału, używane do różnych celów; butelka, dzban, słój, karafka, → alembik.«:

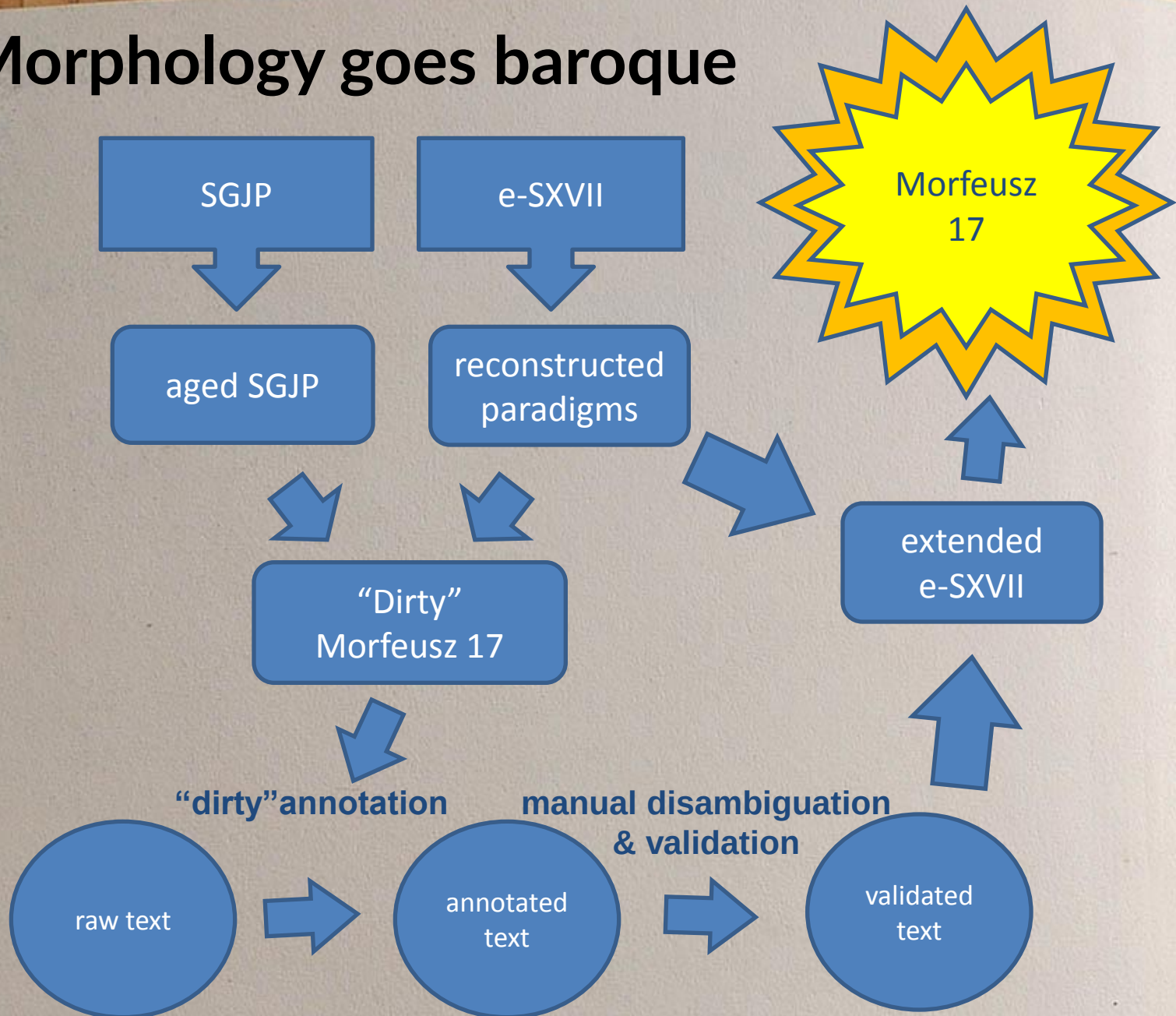
- Zakrystyjan swoją modą Pije wódkę, okrzci wodą, Bańkę z miodem nazwie cieniem I tak - sztucznym omamieniem Mądre głowy oszuka. *MelScholBar II II, 760.*
- Y odpowiedział: Żywie Pan Bog twoj/ żeć nie mam piezonego chlebá/ oprócz z garść pełną mąki w gárcu/ á trochę oliwy w bańce. *BG 1Kr1 17, 12.*
- Tedy rzekł [Elizeusz]: Przynieście mi bańkę nową/ á włożcie w nią soli. *BG 2Kr1 2, 20.*
- A wzięwszy bańkę Olejku/ wylejesz ná głowę jego/ y rzeczesz: Ták mowi PAN: Pomázałem cię zá Krolá nád Jzráelem. *BG 2Kr1 9, 3.*
- Są też dwie bańce trwie ś. Wáwrzyńcá. *WargRzym 80.*

Date of publication: 1610

Morphological processing



Morphology goes baroque



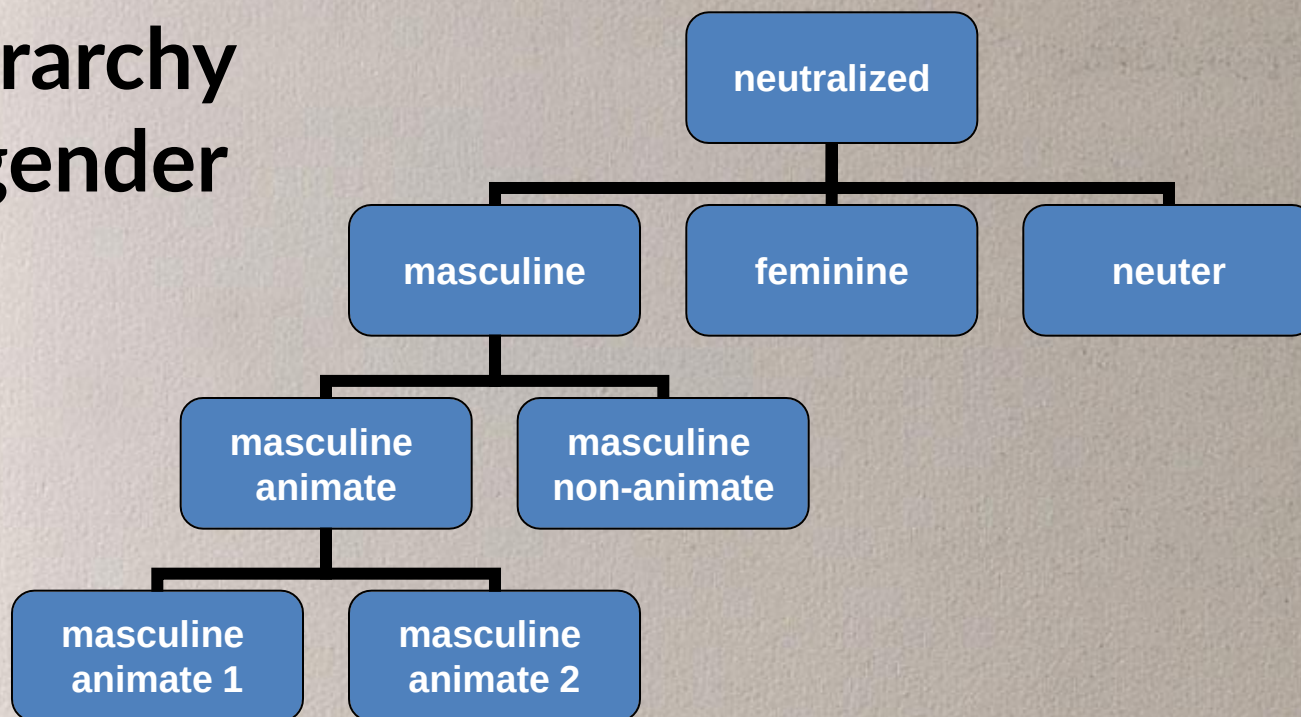
Aging SGJP

- Historical forms added to contemporary paradigms
 - Perfect participle for imperfective verbs (e.g. *widziawszy* 'having seen')
 - Adjectival predicative form (e.g. *słuszna* 'it's proper')
- Additional values of grammatical categories
 - Dual number (e.g. *żabie* '(two) frogs')
 - Past perfect tense (e.g. *widziałem był* 'I had seen')
- New paradigms for regularly created derivatives
 - **Active & passive** participles in comparative (e.g. *bolątszy* 'more hurting')
- Modification within the category of gender

Nominal paradigm supplemented with forms of dual number

	singularis	pluralis	dualis
Nom	komputer	komputery	komputera
Gen	komputera	komputerów	komputeru
Dat	komputerowi	komputerom	komputeroma
Acc	komputer	komputery	komputera
Instr	komputerem	komputerami	komputeroma
Loc	komputerze	komputerach	komputeru
Voc	komputerze	komputery	komputera

Hierarchy of gender



form	gramm. values	lemma	gender
abrysom	Dat. pl	ABRYS (?) ABRYSA (?) ABRYSO (?)	neutralized
abrysem	Instr. sg	ABRYS	masculine
abrys	Acc. sg	ABRYS	masc. non-anim.
abryse	Acc. sg	ABRYSA	feminine

Conversion to transcribed version

Transliterated version	Transcribed version
B. Nie z ^á da mi żaden te ^y rzeczy, kt ^{or} aby b ^{ár} zie ^y owemu niżeli mnie należć nie mi ^á á. A t ^á k śmierć ^y Kupidyn poz ^á rszy się i ^á ko be st ie, posnęli w Koś ^ć iele Bacchusowym, gd ^ź ie niewyszumiawszy z przepi ^ć ia śmierć Cupidin ^{ow} , ^á Cupido śmierci należący przypasawszy s ^á y ^{da} k do boku szły, swych należących odpr ^á wow ^á ć powinnoś ^ć i.	B. Nie z ^a da mi żaden te ^j rzeczy, kt ^ó raby bar ^z ye ^j owemu niżeli mnie należć nie mi ^a á. A t ^a k śmierć ⁱ Kupidyn poz ^a rszy się j ^a ko be ^z tie, posnęli w Koś ^c iele Bacchusowym, gd ^z ie niewyszumiawszy z przepi ^c ia śmierć Cupidyn ^{ów} , ^a Cupido śmierci należący przypasawszy s ^a j ^d ak do boku szły, swych należących odpr ^a wow ^a ć powinnoś ^c i.

Experiment

- A random sample of KORBA was selected – altogether 16.519 tokens from 19 texts
- Transliterated version was converted to transcribed version.
- The sample was analyzed by Morfeusz

Results: 1552 unrecognized forms (9%)

- The same sample was analyzed by Morfeusz supplemented with forms exported from e-SXVII

Results: x unrecognized forms

Comments on the results

- Relatively many forms turned out to be recognizable (Morfeusz can recognize quite a lot of old forms):
 - aliści – old lexeme ALIŚCI ‘however’
 - słowy – old form (Instr. sg) of lexeme SŁOWO ‘word’
- Some forms were unrecognized because of converter’s mistakes:
 - wtym – lack of a segmentation rule (it should be: w tym ‘in that’)
 - puł – lack of a conversion rule (it should be: pół ‘half’)
- Some forms were unrecognized because they have fallen into disuse and are not included in Morfeusz’s paradigms:
 - moje – old form (Acc. sg, f.) of lexeme MÓJ ‘my’

Conclusion

- Main problem with morphosyntactic tagging of a Polish baroque corpus:
 - lack of a baroque grammatical dictionary of Polish, which could be a base for a baroque inflectional analyser
- Ways of facing this problem:
 - joining incomplete data from different sources (grammatical dictionary of contemporary Polish, grammatical forms gathered in a baroque dictionary)
 - inflectional analyser built by successive approximations: preliminary version → manual tagging → extended version

Thank you!