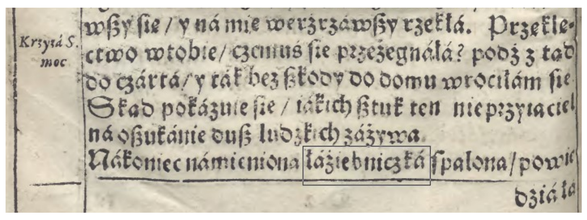
NOTUM PER IGNOTUM

THE ENRICHMENT OF LEXICAL INFORMATION AND CORPUS RESOURCES BY USING THE RESULTS OF MORPHOLOGICAL ANALYSIS OF HISTORICAL TEXTS

THE ELECTRONIC CORPUS OF THE 17TH AND 18TH C. POLISH TEXTS (UP TO 1772) - STEP BY STEP

STEP ONE collecting materials (from a scan to a ready-to-work TEI-tagged text)


Krzyżá S. moc

wßy ſie / y ná mie werzrzawßy rzekłá. Przekleſtwo wtobie / czemuś ſie przeżegnáłá? podż z tad do czártá / y ták bez ſkody do domu wrociłám ſie Skad pokázuie ſie / iákich ſtuk ten nieprzyiaciel ná oßukánie duſz ludzkich zázywa. Nákoniec námieniona łáźiebniczká ſpalona / powie dziáłá

the fragment of the Polish translation of „Mallus Malleficarum" by Kramer and Sprenger – 1614

Nákoniec námieniona łáźiebniczká spalona/ powie

działa

[KONIEC STRONY]

[POCZĄTEK STRONY]

29

Część Pierwsza

działá/ że tákim sposobem od pewney báby byłá zwiedzioná/ áczkolwiek iey towárzyszká

manual transliteration and structural annotation (WORD format)

<note place="margin" xml:id="txt_4.4.3.19-note"> Krzyża S. moc</note>
Przeklęctwo wtobie/ czemuś sie przeżegnáłá? podż z tąd do czártá/ y ták bez szkody do domu wrociłám sie. Stąd pokázuie sie/ iákich sztuk ten nieprzyiaciel ná oszukánie dusz ludzkich zázywa. Nákoniec námieniona łáźiebniczká spalona/ powie
<pb n="29"/>
<fw type="runningHead" xml:id="txt_4.4.3.20-fw">Część Pierwsza</fw>
działá/ że tákim sposobem od pewney báby byłá zwiedzioná/ áczkolwiek iey towárzyszká rozny obyczáiem/ to iest: iż szátáná w osobie ludzkiey podkálá w drodze gdy vmysłem płodzenia wszeteczeństwá/ szłá miłośniká swe^o^ náwiedzác: z ktor obyczáiem/ to iest: iż szátáná w osobie ludzkiey podkálá w drodze/ gdy vmysłem płodzenia wszeteczeństwá/ szłá miłośniká swe^o^ náwiedzác: z ktorý szátánem obyczáiem ludzkim cieleśnie złączywszy się/ spytał iey szátan/ iesliby go poznáłá/ odpowiedziálá oná/ że go by namniey nie zna: odpowiedział on. Iestem szátan/ y iesli

convertion to XML (compatible with TEI format)

The highlighted form is *łaziebniczka* 'a woman serving in baths' – a word that haven't been recognised before. Let's see the path that the string should go through to be recognised.

łaziennik *m III przestarz.* in. łaziebnik: Łaziennik wyciągał członki, szorował skórę, wycierał, a na ostatek namaszczał ciało olejami wonnymi z mirrą, nardem, cynamonem. KRASZ. *Kartki 669. // L*

masculine form of „łaziebniczka" from Doroszewski's dictionary – 1958

STEP TWO from a transliterated text to a transcribed and morphosyntactically tagged text

AMEBA SUPERTOOL (from Pol. *ameba* 'amoeba', Gr. *ἀμοιβή* 'change')

a tool converting texts from their transliterated to transcribed version

TRANSKRYBER

formulas of transcription for the converter orginally created by Janusz Bień and his team for the purposes of the IMPACT project

– https://bitbucket.org/jsbien/pol

id_1768	1 EM	^za	mow	i*$	mów
id_1769	1 EM	^za	mowi	T.*	mówi
id_1770	1 EM	^za	mowic	$	mówić
id_1771	1 EM	^za	mowi	Ą.*	mówi
id_1772	1 EM	^za	mowić	$	mówić
id_1773	1 EM	^u	mow	i*$	mów
id_1774	1 EM	^u	mowi	T.*	mówi
id_1775	1 EM	^u	mowic	$	mówić
id_1776	1 EM	^u	mowi	Ą.*	mówi
id_1777	1 EM	^u	mowić	$	mówić
id_1778	1 EM	^wy	mow	i*$	mów
id_1779	1 EM	^wy	mowi	T.*	mówi
id_1780	1 EM	^wy	mowic	$	mówić
id_1781	1 EM	^wy	mowi	Ą.*	mówi
id_1782	1 EM	^wy	mowić	$	mówić
id_1783	1 EM	^o	mow	i*$	mów
id_1784	1 EM	^o	mowi	T.*	mówi
id_1785	1 EM	^o	mowic	$	mówić

KORBEUSZ

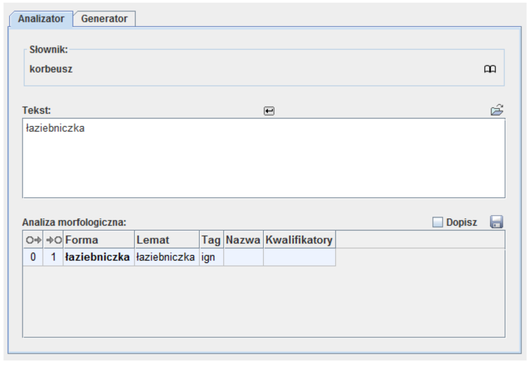
At this point, every string should be converted to the modern form. The diacritical marks such as acutes or graves are erased (if not in use nowadays) or are added (if the opposite situation occurs). The orthography is modernised (contemporary rules of orthography are applied).

Special care should be taken here; if the formula is not restricted enough (especially by adding exceptions), it is extremely easy to deform a correctly written string.

Every move can throw the baby out with the bathwater.

morphological analyser customised to analyse 17th and 18th C. texts, based on Morfeusz – an analyser for contemporary Polish.

http://sgjp.pl/morfeusz/



Two main resources of the Korbeusz analyser are (the information are harvested from):

THE ELECTRONIC DICTIONARY OF THE 17TH-18TH C. POLISH: E-SXVIII

partially prepared inflectional information about lexemes

http://xvii-wiek.ijp-pan.krakow.pl/pan_klient/

THE GRAMMATICAL DICTIONARY OF POLISH: SGJP

set of the paradigms of contemporary Polish

http://sgjp.pl/

Lack of the information in one of the resources causes the inability to recognise the string, delimitate, and then lemmatize it and give it the adequate values of grammatical categories. Such forms are harvested and compiled (as a list of unrecognised strings). And now the real work begins; sifting through the given data and evaluating it.

The fragment of the ign-list below shows the discussed string.

246466	SpInZąbMłot	zwiędziony	zwiędziony
246467	SpInZąbMłot	zlę	zlę
246468	SpInZąbMłot	żebymći	żebymci
246469	SpInZąbMłot	sercecznie	sercecznie
246470	SpInZąbMłot	Bázylius	Bazylius
246471	SpInZąbMłot	nowu	nowu
246472	SpInZąbMłot	świątobliwsze	świątobliwsze
246473	SpInZąbMłot	pociągńáć	pociągnąć
246474	SpInZąbMłot	Ráwenspurgu	Rawenspurgu
246475	SpInZąbMłot	záchosuią	zachosują
246476	SpInZąbMłot	łáźiebniczká	łaziebniczka
246477	SpInZąbMłot	wazyłá	wazyła
246478	SpInZąbMłot	obywátelŏ	obywatelŏ
246479	SpInZąbMłot	gorniego	górniego

working on *igns* – seeking the balance between analysis paralysis and corpus utopia

the ign-lists (a set of strings unrecognised by Korbeusz and tagged as *ign* – from Lat. *ignotum* 'unknown')

	A	B	C
1	text	translit	orth
2	DolBitKontr	MATYJASZ	MATYJASZ
3	DolBitKontr	JUNII	JUNII
4	DolBitKontr	JUNII	JUNII
5	DolBitKontr	MATYJASZA	MATYJASZA
6	DolBitKontr	SUPRA	SUPRA
7	DolBitKontr	PENDENTA	PENDENTA
8	DolBitKontr	JUNII	JUNII
9	DolBitKontr	wielłebności	wielłebności
10	DolBitKontr	senijorów	senijorów

currently: ca. 0.5M segments

	A	B
1		
2	orth	Liczba - orth
3	jm	4172
4	jmp	2898
5	Zaczym	2609
6	beł	1595
7	Cezarza	1558
8	Cezarz	1345
9	Wm	1205
10	ichm	1001

ca. 0.2M unique segments (not lemmas)

Evaluating the igns is the key to enrich our resources successfully. The main factor is frequency. More frequent strings should be added to dictionary first to decrease the amount of unrecognised strings. It is also a good time to check out if the transcription is applied correctly.

STEP THREE, FOUR, FIVE ...

enrichment of lexical information and corpus resources

1. adding new entries or new inflectional forms to the e-SXVIII


POLSKA AKADEMIA NAUK | INSTYTUT JĘZYKA POLSKIEGO
ELEKTRONICZNY SŁOWNIK JĘZYKA POLSKIEGO XVII I XVIII WIEKU

| o słowniku | lista haseł | zaawansowane wyszukiwanie | kwerendy |

ŁAZIEBNICZKA rzecz. ż

Najwcześniejsze poświadczenie: 1614

Formy gramatyczne:

lp

M łaziebniczka

»kobieta posługująca w łaźni«:

– Nákoniec námieniona łáźiebniczká spalona/ powiedziáłá/ że tákim sposobem od pewney báby byłá zwiedzioná [...]. *SpInZąbMłot 28-29.*

2. creating new formulas for Transkryber

To make the string recognizable by Korbeusz it is necessary to modernise the orthography. In our case, at least two formulas should be used:

~ *if „ź" stands before „i", it should be replaced with „z"*

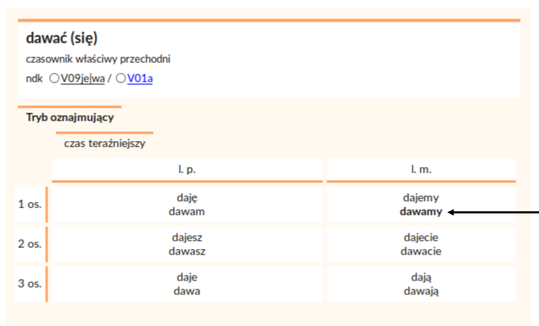
~ *every „á" should be replaced with „a"*

So it goes like the following:

id_of_formula	on/off	inicials_of_author	right_context	replacee	left_context	replacer
id_1	1	EM	.*	ź	i.*	z
id_2	1	EM	.*	á	.*	a

input: łáźiebniczká output: łaziebniczka

3. adding new paradigms to the SGJP


dawać (się)

czasownik właściwy przechodni

ndk ○V09jelwa / ○V01a

Tryb oznajmujący

czas teraźniejszy

	l. p.	l. m.
1 os.	daję dawam	dajemy **dawamy**
2 os.	dajesz dawasz	dajecie dawacie
3 os.	daje dawa	dają dawają

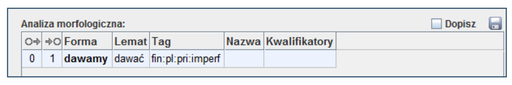
Sprawy dawne z duchownemi o dziesięciny aby były kasowane, chociażby były osądzone i na egzekucyej były, a teraz znowu dla zgody z nimi, jabym rozmiał in genere wszystkie, począć dawać według postanowienia; rozumiałbym, żebyśmy tym tylko płacili, którzy w naszych parafiach służą i robią i dawać jem tak, coby się dobrze wychowali, służąc Panu Bogu, a jeśliby co postronnym dalekim przyszło dawać, tedy tak dawać z łanu, jako pobór **dawamy,** i tylko w ten czas, kiedy poboru niemasz, gdyż mają wsi i wielkie opatrzenia. A w każdej sprawie in causis simplicibus niechby był każdy bliższy do odwodu tak z duchownych, jako świeckich. Bo teraz duchowni świeckich każą konwinkować, coby nie miało być; citatus propior być ma in causis simplicibus ad evasionem . Gdy granice arcybiskup, biskup, opat i każdy z duchownych z którym ślachcicem odprawuje, niech też sam

„dawamy" is an old form of the lexeme DAWAĆ 'to give' (plural form of the first person), was added to the SGJP as a variant of modern form „dajemy"

and then, retrieving information by Korbeusz


Analiza morfologiczna: Dopisz

		Forma	Lemat	Tag	Nazwa	Kwalifikatory
0	1	**łaziebniczka**	łaziebniczka	subst:sg:nom:f	*nazwa pospolita*	


Analiza morfologiczna: Dopisz

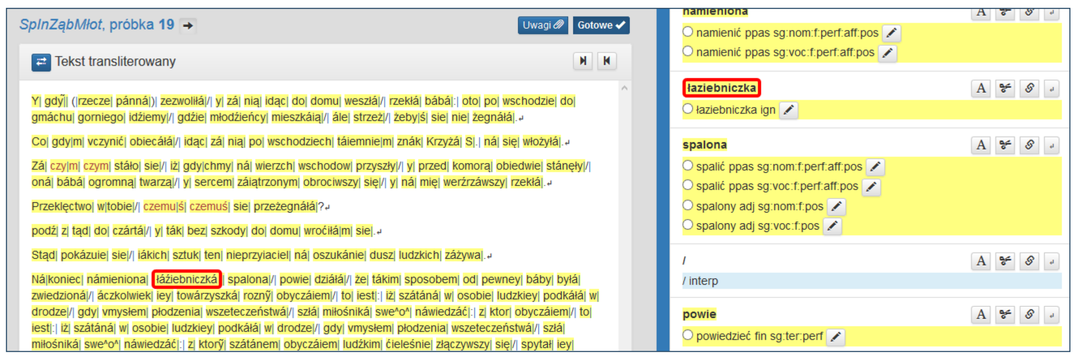
		Forma	Lemat	Tag	Nazwa	Kwalifikatory
0	1	**dawamy**	dawać	fin:pl:pri:imperf		

... AND ALMOST OUR GOAL

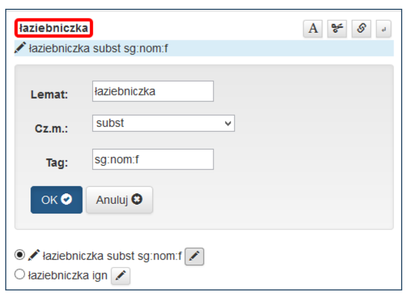
manual disambiguation and creating the corpus

ANOTATORNIA

a tool supporting manual morphosyntactic annotation of texts; the annotators disambiguate interpretation of the segment by selecting the one provided beforehand by Korbeusz or adding their own manually (as it is shown below).



The discussed word (łaziebniczka) is still unrecognised by Korbeusz due to the lack of harvesting. In such case, the adequate values of grammatical categories can be chosen manually by typing a lemma and selecting corresponding POS and tags.



Jedna	łaziebniczka [::::]	/ między inszemi rzeczami do których się przyznała to też	
wróciła/ grad przyszedł. I tak trzeciego dnia były spalone: a	łaziebniczka [::::]	załując za grzechy swoje/ i wyspowiedawszy się ich/	
ale coby czyniła niewiem. A nakoniec spytana	łaziebniczka [::::]	/ jakoby czas wielki wyszedł/ od wzięcia wody aż	
aczkolwiek i inszych było niemało/ imię jednej było Agnieszka	łaziebniczka [::::]	/ a drugiej Anna z Mindelhein które pojmawszy dano do	
Nakoniec namieniona	łaziebniczka [::::]	spalona/ powie działa/ że takim sposobem od pewnej	

The very test version of KORBA corpus

ABOUT OUR PROJECT:

title:	Electronic corpus of 17th and 18th century Polish texts (up to 1772), acronym: KorBa (KORpus BArokowy)
the aim:	the creation of a corpus of 17th and 18th century Polish texts (another stage of development of the Polish National Corpus (Narodowy Korpus Języka Polskiego, NKJP, see: http://nkjp.pl/) and tools for its processing (searching, filtering, summarizing statistical data, etc.)
corpus size:	around 12 million tokens
duration:	2013 – 2018
coordinator:	Institute of Polish Language, Polish Academy of Sciences
cooperation:	Linguistic Engineering Group, Institute of Computer Science, Polish Academy of Sciences
main investigator:	Włodzimierz Gruszczyński
funding:	A Ministry of Science and Higher Education National Programme for the Development of Humanities grant 0036/NPRH2/H11/81/2012

CONTACT US:

Renata Bronikowska
r.bronikowska@wp.pl

Emanuel Modrzejewski
modrzejewski.emanuel@gmail.com

Institute of Polish Language, Polish Academy of Sciences
www.ijp-pan.krakow.pl/en

The Electronic Corpus of the 17th and 18th c. Polish Texts (up to 1772)
korba@ijp-pan.krakow.pl

Lato Fonts designed by Łukasz Dziedzic