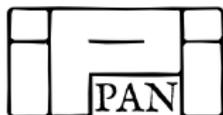


Basic natural language processing toolkit for 19th century Polish

Witold Kieraś Marcin Woliński

Linguistic Engineering Group
Institute of Computer Science
Polish Academy of Sciences



September 26th 2018, Prague

❧ Introduction ❧

- ➡ Digitisation of library archives brings huge amounts of data which can be used in linguistic research.
- ➡ Computational and corpus linguists are getting more and more interested in historical linguistics bringing their methodology to the field.
- ➡ However digitisation is merely a first step in building historical corpora. Computational tools and resources need to be extended and adjusted to suit the specifics of historical languages.
- ➡ The presentation focuses on NLP tools for 19th c. Polish, however they were developed in close cooperation with the Baroque Corpus of Polish team.

Historical text processing toolkit

-  Small gold-standard training corpus,
-  Automatic transcriber,
-  Morphological analyser (based on Morfeusz 2 SGJP),
-  Anotatornia 2 – web-based system for manual annotation of historical texts,
-  Morphosyntactic tagger.

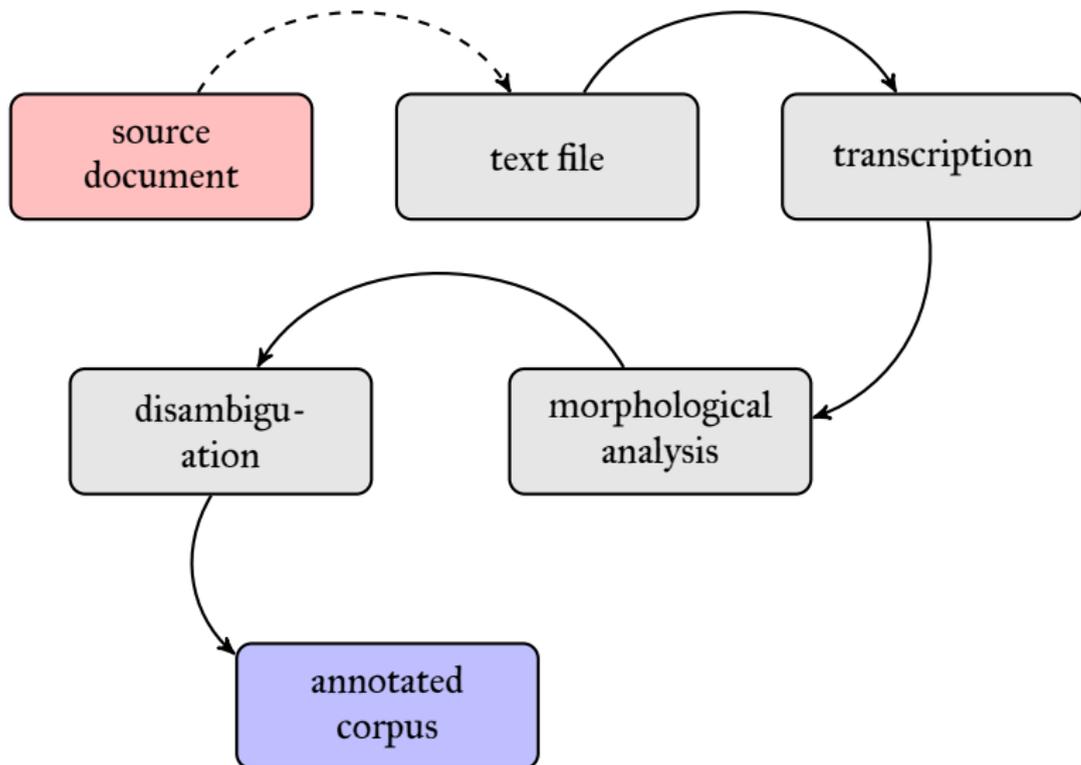
🌀 The Corpus of 1830-1918 I 🌀

- 👉 Collection of samples excerpted from Polish texts.
- 👉 Consists of 1000 samples of ca. 1000 words each = over 1 million words large.
- 👉 Divided into five subcorpora of equal size representing: fiction, essays, science and popular science, short newspaper articles, and drama.
- 👉 Evenly distributed between years: for every year and every stylistic subcorpus there is at least one and at most four samples.
- 👉 Represents all major Polish literary centres in all five stylistic subcorpora, however a bias towards the capital city is significant (nearly 40%) comparing to other major publishing centres: Lviv (16%), Cracow (12%), Poznań (7%) and Vilnius (5%).

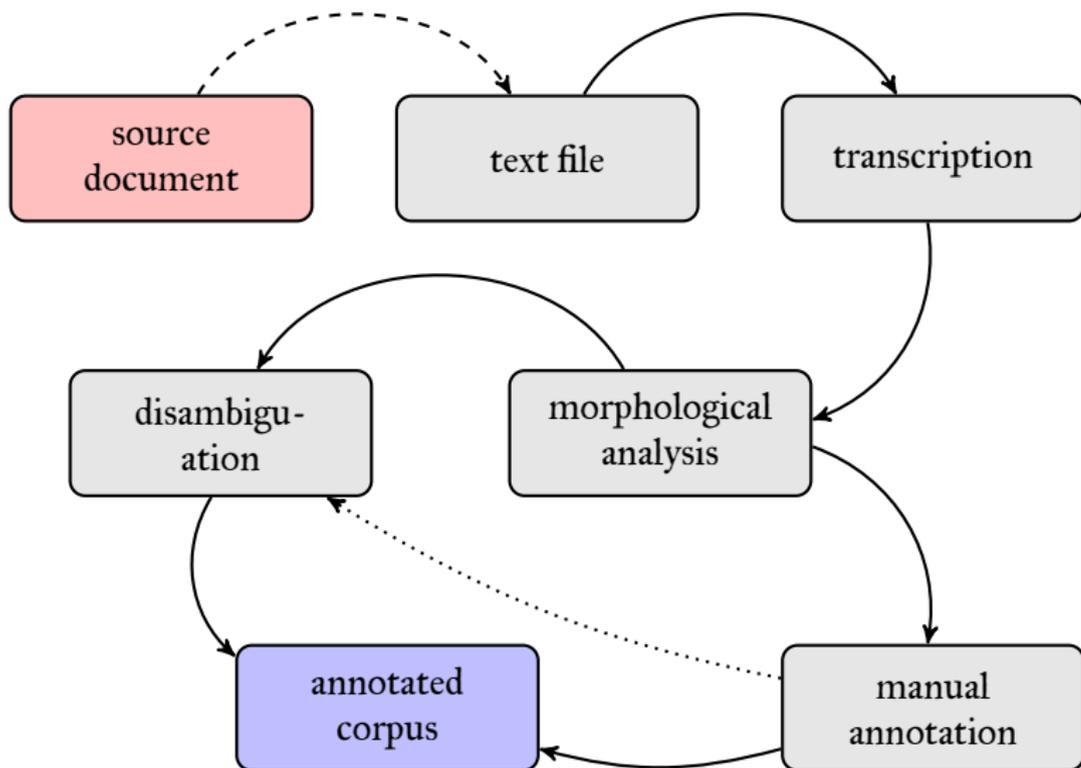
✿ The Corpus of 1830-1918 II ✿

- ✿ The corpus was collected in one of the previous projects conducted at the University of Warsaw (2013-2017, head: Magdalena Derwojedowa).
- ✿ For manual annotation we excerpted 2944 shorter samples of ca. 160 words each, which means that from each original sample three smaller samples were excerpted for manual annotation.
- ✿ The samples needed to be automatically transcribed and preannotated.

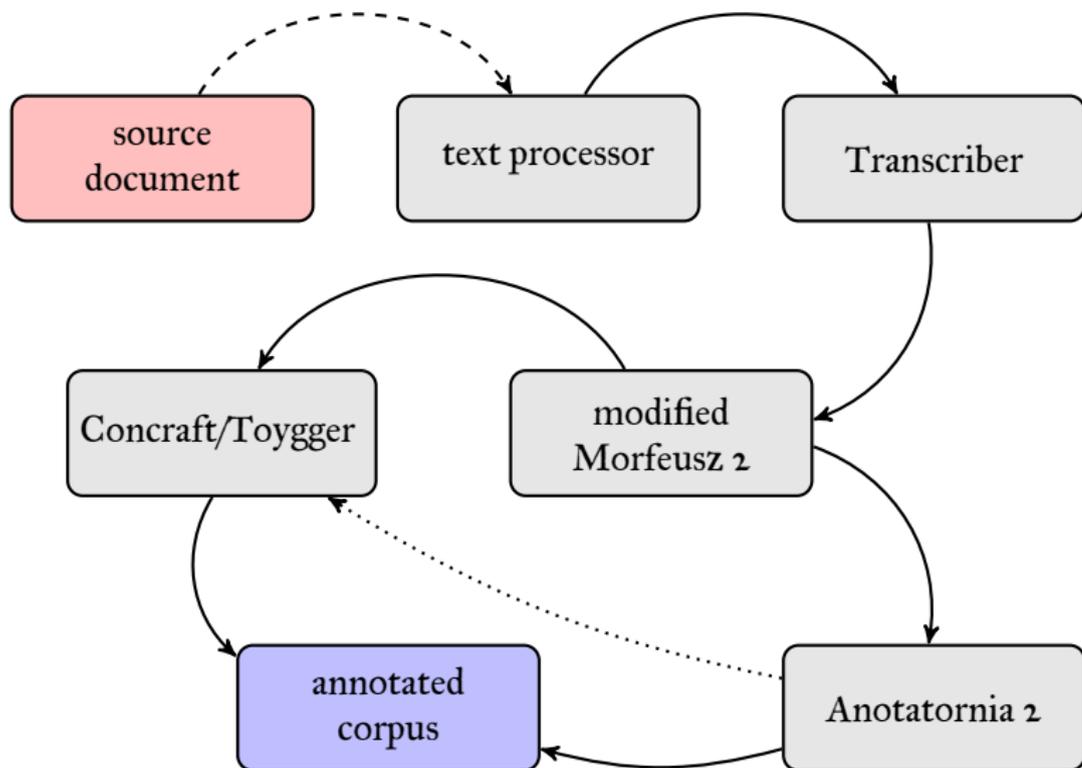
Workflow scheme



Workflow scheme



Workflow scheme



Transcription

- ✎ Normalization concerns only spelling variants, old inflectional forms are not modified to adhere to contemporary language.
- ✎ Thus contemporary lexical data used in NLP tools needs to be modified to catch the inflectional phenomena.
- ✎ It is not always clear how to draw a line between spelling and inflectional phenomena.

Transcription – example

KOMISJA (between 1650 and 1918):

commissiey, commissyi, komisja, komisją, komisję, komisji, komissja, komissya, komissyę, komissyi, komisya, komisją, komisye, komisye, komisyi, komisyj, komisją, komisję, kommissij, kommissja, kommissjów, kommissii, kommissij, kommissją, kommissje, kommissjów, kommissya, kommissyi, kommissyiey, kommissyi.

In the process of transcription:

-  initial *c* was changed to *k*,
-  doubled *s* and *m* were changed to single,
-  letter *j* introduced instead of *y* or *i*.
-  archaic inflectional endings left unchanged: *komiszej, komisjów* and *komisją* (in sg:acc.).

Modified Morfeusz analyser

Dictionary:

- 👉 “Aged” Grammatical Dictionary of Polish data — tagset conversion and reconstruction of regular archaic forms (*talenta, ładną, ładnemi, bogactwy, Chłirczykowie* etc.),
- 👉 Fortunately GDP contains the whole vocabulary of Doroszewski’s dictionary (1958-1969), the lexical database of which dates back to the last quarter of 18th.

Segmentation rules:

- 👉 joint and disjoint spelling of *nie* particle,
- 👉 joint spelling of various clitics: (-ć, -li, -że, -ż),
- 👉 auxiliary enclitic BYĆ ‘to be’ much more mobile comparing to contemporary Polish.

- 👉 Only 1.72% of tokens in our corpus did not receive any interpretation from the analyser.

✿ Taggers ✿



Concraft:

- * conditional random fields,
- * resolves ambiguous segmentation (f-score: 95.85%),
- * morphosyntactic disambiguation accuracy: 90.8%,
- * slightly better guessing.



Toygger:

- * recurrent neural networks (bi-LSTM) with word embeddings,
- * performs only morphosyntactic disambiguation based on a given segmentation,
- * better disambiguation results (93.94%),

Both used simultaneously in automatically annotated corpora as separate layers of annotation, which allows for a more detailed qualitative evaluation of the taggers.

Anotatornia 2

 Anotatornia 2 — web-based system aimed at manual morphosyntactic annotation and verification of corpora.

 Geared towards historical corpora:

- * presents two parallel text forms: transliterated and transcribed
- * preserves information about page number from the original editions for every token.

<http://zil.ipipan.waw.pl/Anotatornia2>

🌀 Anotatornia 2: annotation mode AT+A 🌀

The regular annotation process involves two human annotators working on the same sample independently. Collisions between the two annotators are resolved by the adjudicator.

Time and work saving alternative:

- 👉 Each sample is annotated by only one human annotator.
- 👉 The annotator's work is confronted with an automatic tagger.
- 👉 The collisions between human annotator and the tagger are revised and resolved by another annotator (adjudicator).

← PotFrasz2Kuk_II, próbka 1

Uwagi

Gotowe

Tekst transliterowany

jeden| grzech|,| lecz| każdy| inszy| we| mnie| czyta|.↵

Nie| człek|,| ale| natura| człeczka| tu| wykroczy|,| Na| świat|,|
nie| w| się| mu| w| głowie| obróciwszy| oczy|.↵Jeśli| jakie| przezwiska|,| takie| i| humory|,| Morstyn| sobie|
Oborska| przybiera| do| sfory|.↵Nie| wzdryga| się|,| imiona| przezwiska| łagodzą|:
Stanisław| z| Konstancją| do| tej| ligi| wchodzą|.↵Pytam|,| mój| miły| bracie|,| co|ć| też| było| po| tem|,|
Niepotrzebnym| się| **cale** zarażać| kłopotem|?↵Albo|ś| szalał|,| albo|ś| się| natenczas| był| upił|,| Kiedy|ś|
Kiedy|ś| szyderstwo| z| siebie| w| tym| urzędzie| kupił|.↵Na| kiego|ż| kata| w| Polsce| pieniądze| nie| tracić|,| Gdy|
trzeba|,| choć| kto| błaznem| chce| zostać|,| zapłacić|?↵Leśne| dryjady| i| wy|,| morskie| wiedzcie| nimfy|,| Że| ten|
pan| dzisiaj| został| kpem| za| swoje| tyńfy|.↵

Niepotrzebnym

niepotrzebny
adj sg:inst:m:pos

się

się qub

cale

- cal subst pl:nom:manim2
- cal subst pl:nom:m
- cal subst pl:acc:manim2
- cal subst pl:acc:m
- cal subst pl:voc:manim2
- cal subst pl:voc:m
- cale adv

zarażać

- zarażać inf imperf

kłopotem

- kłopot subst sg:inst:m

?

Annotation results

- ✎ The annotators generated 14.27% conflict rate with the tagger.
- ✎ As expected, large majority of the conflicts were resolved in favour of human annotators (87.22%) however a significant number of human errors were also found and corrected as the remaining 12.78% have been either resolved in favour of the tagger (6.69%) or changed to an alternative interpretation provided by the adjudicator (6.09%).
- ✎ As a result of the annotation process, 2944 samples were annotated by one human annotator and tagger. The total number of 625,000 tokens were annotated in the project.

Corpora

- ➡ Manually annotated gold-standard subcorpus of 1830-1918 (0.6M),
- ➡ Automatically annotated corpus of 1830-1918 (1.3M),
- ➡ Automatically annotated corpus of 102 novels by Józef Ignacy Kraszewski published before 1888, downloaded from Wikisource.pl (8.2M).

<http://korpust9.nlp.ipipan.waw.pl/>

KORPUS TEKSTÓW POLSKICH Z LAT 1830-1918



Korpus

Korpus automatyczny (1,3 mln)

Zapytanie

[tag="adj;sg:acc:f:." & orth=".*ę" & lbase="ten"]

á

é

Á

É

ZAAWANSOWANE ▾

KONSTRUKTOR ZAPYTAŃ

Wyszukaj

Znaleziono 140 wyników.

Lp	Lewy kontekst	Rezultat	Prawy kontekst	Etykieta	Data
1	. a wiesz ty kto jest ta biedna dziewczyna,	którę [które;adj;sg:acc:f:pos]	zranił mój Jasio zamiast syna twego, którą twój syn	1847_4.2	1847
2	dziewczyna, którą zranił mój Jasio zamiast syna twego,	którę [które;adj;sg:acc:f:pos]	twój syn zwiódł i shańbił.... czy	1847_4.2	1847
3	pytam się ciebie. Za to, że zdradza tak	szlachetnę [szlachetne;adj;sg:acc:f:pos]	duszę, jaką ty masz, siostrzo? Nigdy?	1910_5.2	1910
4	Do syna po odejściu Pszonki). Ha! na	moję [mój;adj;sg:acc:f:pos]	duszę, Przebrałeś miarę, synku, udając ciemnę	1861_5.1	1861
5	oczu ludzie nie widzieli. Cześniak. Korczyż więc	naszę [nasz;adj;sg:acc:f:pos]	sprawę jakąśmy poczęli: Niech się wreszcie nad nami	1861_5.1	1861
6	, to jest sztuka: Przez nie, może cześć	naszę [nasz;adj;sg:acc:f:pos]	podać wiek wiekowi. Nam wolna śmiać się, bredzić	1861_5.1	1861
7	, że musi rzucić się na bok aby szcękami ująć	swoję [swój;adj;sg:acc:f:pos]	zdobycz, jakże więc mógłby ten manewr wykonać nie	1840_3.1	1840 (3 kwietnia)
8	było tak, jak żądam. Hrabia Toż mi zamęczysz	moję [mój;adj;sg:acc:f:pos]	biedną żonę. Bo powiem ci, pod sekretem przyznała	1845_5.2	1845
9	tak mało. A wartoby, żebyście ubiegającą	naszę [nasz;adj;sg:acc:f:pos]	teraźniejszość utrwalił na płótnie, za przykładem braci waszych poetów	1845_5.2	1845

KORPUS TEKSTÓW POLSKICH Z LAT 1830-1918



Korpus

Korpus automatyczny (1,3 mln)

Zapytanie

[tag="adj;sg:acc:f:." & orth=".*ę" & lbase="ten"]

á

é

Á

É

ZAAWANSOWANE ▾

KONSTRUKTOR ZAPYTAŃ

Metadane

Etykieta ▾

Ograniczenie

zaczyna się od ▾

Zapytanie o metadane

 Statystyki

Grupowanie wg

Data ▾

Długość przedziału

1 rok ▾

Liczba wyników na stronę

20 ▾

Warstwa wyświetlania

transliterowana ▾

Wyszukaj

Znaleziono 140 wyników.

Lp	Lewy kontekst	Rezultat	Prawy kontekst	Etykieta	Data
1	. a wiesz ty kto jest ta biedna dziewczyna,	którę [które:adj;sg:acc:f:pos]	zranił mój Jasio zamiast syna twego, którą twój syn	1847_4.2	1847
2	dziewczyna, którą zranił mój Jasio zamiast syna twego,	którę [które:adj;sg:acc:f:pos]	twój syn zwiódł i shańbił.... czy	1847_4.2	1847
3	pytam się ciebie. Za to, że zdradza tak	szlachetnę [szlachetne:adj;sg:acc:f:pos]	duszę, jaką ty masz, siostrze? Nigdy?	1910_5.2	1910

19	nie zginie w czarnej norze Teni, co w pomoc	Twoje [twój:adj:sg:acc:f:pos]	wierzy, Z piekiet ratuj go obieży. (Słychać	1890_5.1	1890
20	uda mu się tak, jak przeszłego roku, zabrać	moje [mój:adj:sg:acc:f:pos]	pszenicę za pół darmo... Nie doczekanie!	1892_4.1	1892

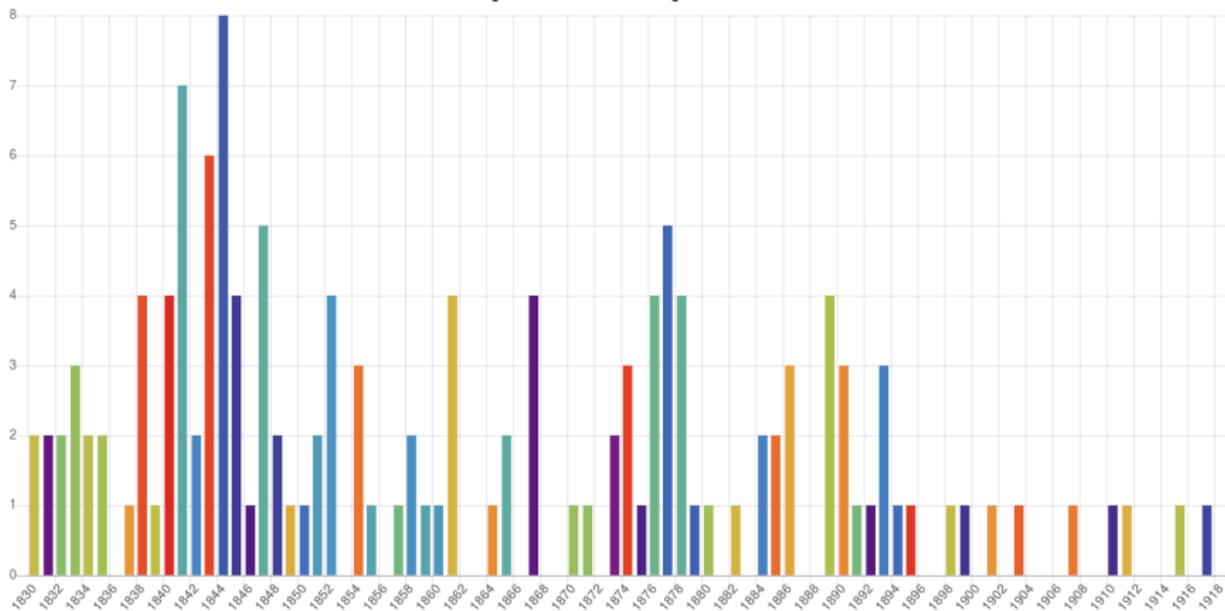
Pobierz wyniki (CSV)



OPIS JĘZYKA ZAPYTAŃ



WYSTĄPIENIA ZE WZGLĘDU NA: DATA



KORPUS TEKSTÓW POLSKICH Z LAT 1830-1918



Korpus

Korpus automatyczny (1,3 mln)

Zapytanie

[base="komisja" & translit="."*mm.*"]

á

é

Á

É

ZAAWANSOWANE ▾

KONSTRUKTOR ZAPYTAŃ

Wyszukaj

Znaleziono 64 wyników.

Lp	Lewy kontekst	Rezultat	Prawy kontekst	Etykieta	Data
1	niniejszego Postanowienia w Dzienniku Praw, Dyrektorowi Głównemu Prezydującemu w	Komisji [komisja:subst:sg:loc:f]	Rządowej Sprawiedliwości poleca się. W Warszawie, d.	1842_2.2	1842 (29 września)
2	(podpisano) Xiąże Warszawski. Dyrektor Główny Prezydujący w	Komisji [komisja:subst:sg:loc:f]	Rządowej Sprawiedliwości. (podpisano) Kossecki. (Gaz	1842_2.2	1842 (29 września)
3	, publika i prasa dobrze przyjęły ten układ, a	komisja [komisja:subst:sg:nom:f]	Izby wybrana z członków większości popierającej ministerjum gotową była popierać	1868_2.1	1868 (25 lipca)
4	popierać zatwierdzenie Tymczasem wcale niespodzianie ta	komisja [komisja:subst:sg:nom:f]	zaproponowała zupełną zmianę konwencji zawartej między ministrem skarbu i towarzystwem	1868_2.1	1868 (25 lipca)
5	ż nadzieję, że Izba się namyśli i pomimo propozycji	komisji [komisja:subst:sg:gen:f]	, zatwierdzi układ już zawarty. Sprzedaż dóbr kościelnych odbywa	1868_2.1	1868 (25 lipca)
6	roku zaciągną do wojska tylko 40,000 młodych ludzi, chociaż	komisja [komisja:subst:sg:nom:f]	Izby deputowanych proponowała rządowi zaciąg 50,000 ludzi. Demonstracje i	1868_2.1	1868 (25 lipca)
7	Petersburski pisze co następuje: w Listopadzie zeszłego roku do	Komisyyi [komisja:subst:sg:gen:f]	przyjmującej rekrutów w Niżnym-Nowgorodzie, przecisnął się włościanin Xięcia Repnina	1834_2.3	1834 (16 marca)
8	brata jego. Mimo cały interes jaki wzbudził w	Komisyyi	nie mogli oni usku...ecznic szlachetnej jego	1834_2.3	1834 (16 marca)

🌀 Conclusions 🌀

- 👉 We have built and tested a production path from raw text data to morphosyntactically annotated searchable corpus.
- 👉 It was implemented in two projects concerning historical corpora.
- 👉 There is much more work to be done in the area. Chances are that some other historical projects will receive financial support and will reuse the technological path.
- 👉 One way of improving the toolkit is to train a machine learning system for automatic transcription as we now have the manually prepared gold data.
- 👉 The reported work was partially supported by a National Science Centre, Poland grant DEC-2014/15/B/HS2/03119.
Title: A diachronic formal model of Polish inflection and its implementation
Time span: August 2015 – February 2019

Transcription rules – example

	left context	match	right context	change to	exceptions
1.	.*	ôô	.*	go	–
2.	^	naiw	.*	najw	naiwn.*
3.	T	j	T	y	mjr

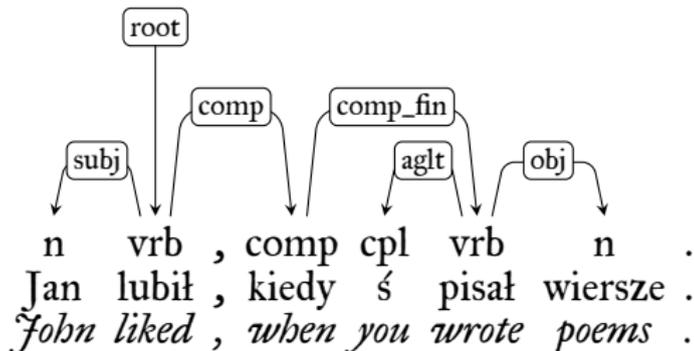
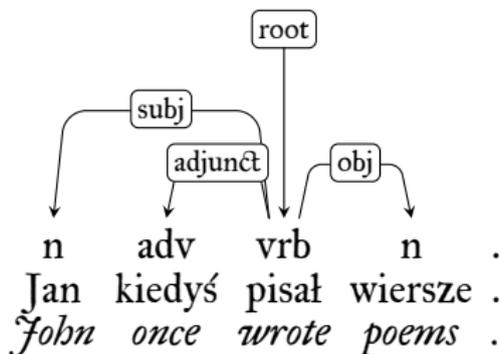
❧ Ambiguous segmentation ❧

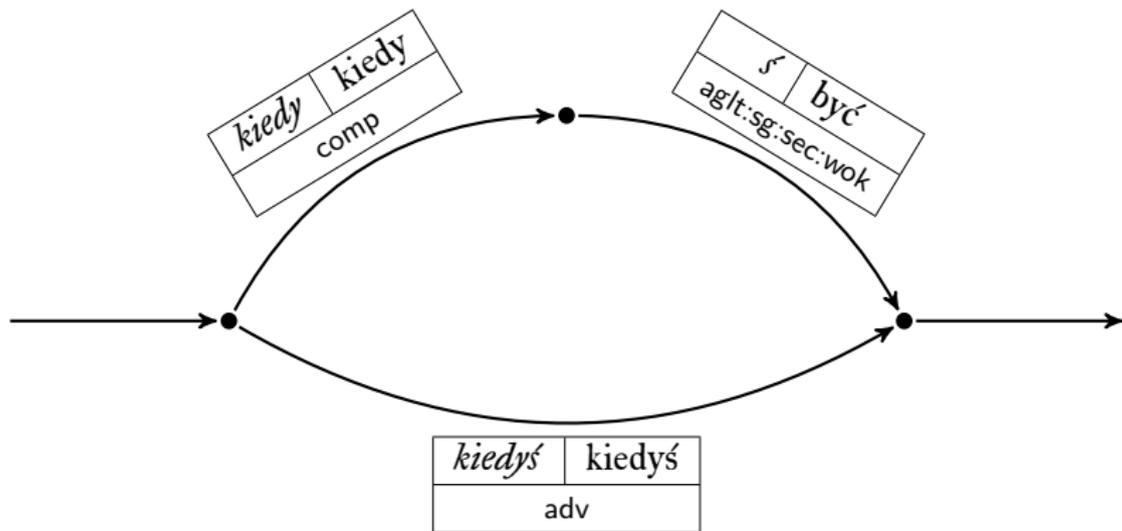
Jan kiedyś pisał wiersze.

Jan lubił, kiedyś pisał wiersze.

- (1) Jan kiedyś pisał wiersze.
John once wrote poems.
- (2) Jan lubił, kiedyś pisał wiersze.
John liked when you wrote poems.

❧ Ambiguous segmentation ❧





Rysunek : Directed acyclic graph (DAG) analysis of the word *kiedys*