Uniwersytet
Wrocławski

Digital humanities in literary studies part 2.

dr hab. Maja Pawłowska

CZ PL
PŘEKRAČUJEME HRANICE
PRZEKRACZAMY GRANICE
2014—2020

EVROPSKÁ UNIE / UNIA EUROPEJSKA
EVROPSKÝ FOND PRO REGIONÁLNÍ ROZVOJ
EUROPEJSKI FUNDUSZ ROZWOJU REGIONALNEGO

Uniwersytet Wrocławski i Uniwersytet Palackiego w Ołomuńcu wspólnie realizują projekt
współfinansowany przez Unię Europejską

HUMANISTYKA CYFROWA DLA PRZYSZŁOŚCI

Areas of digital literary studies

a. digital literature

b. digital sources of literature

c. digital applications for text and document analysis

d. digital methods in (semi-)automatic translation

e. digital text editing

f. automatic taxonomies

Digital text analysis

Digital text analysis

The area of digital analysis of literary (artistic) texts includes:

– determining the filiation of texts;
– examining the authorship of texts;
– creating numerical taxonomies of texts.

In addition, automatic overtone analysis (sentiment analysis) can be conducted in such texts.

Digital text analysis

Text filiation is a technique known since at least the 19th century.

Its object is to link multiple sources of a text into a network of relationships.

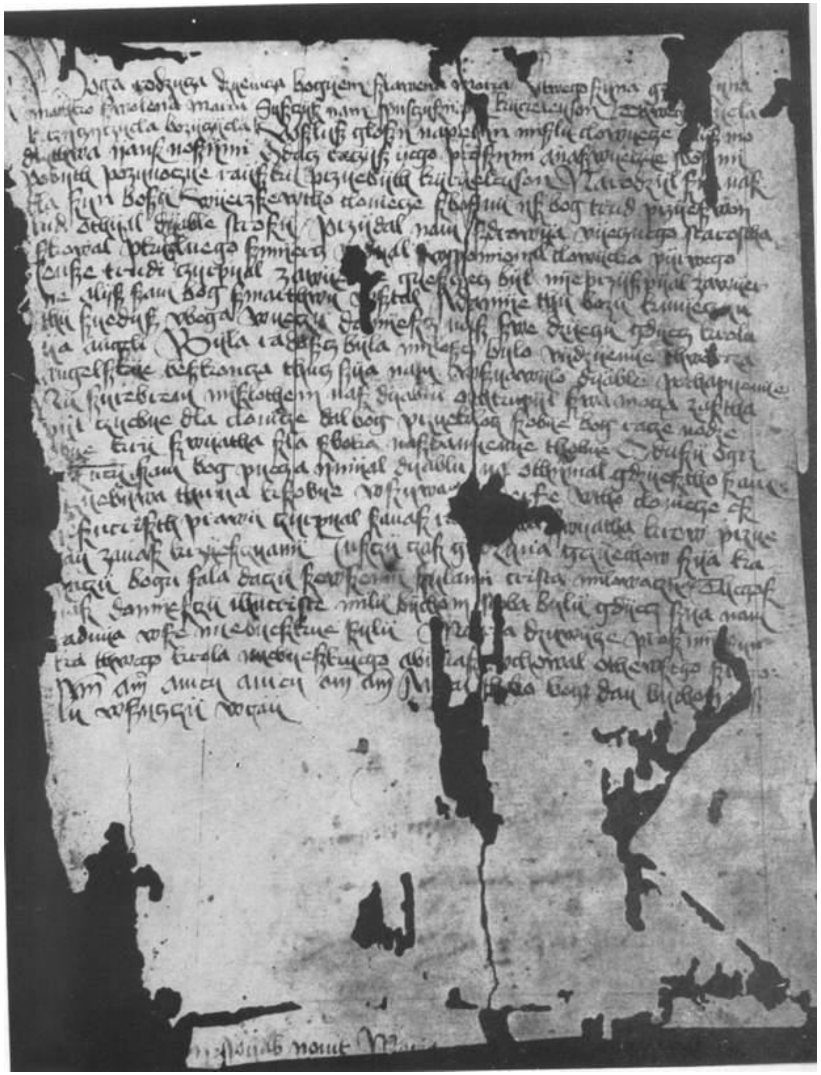
This mostly applies to texts from bygone eras, which did not leave a sufficient number of records because oral communication was dominant.

It is about scattered fragments of records of a work of oral literature, which contain its earlier versions, "leading" to the final version of the work.

Manuscript of the Bogurodzica from Codex C 423.

The song as we know it today consists of three parts, created at different times.

Filiation consists in indicating their relationship.



Authorship studies and digital taxonomies

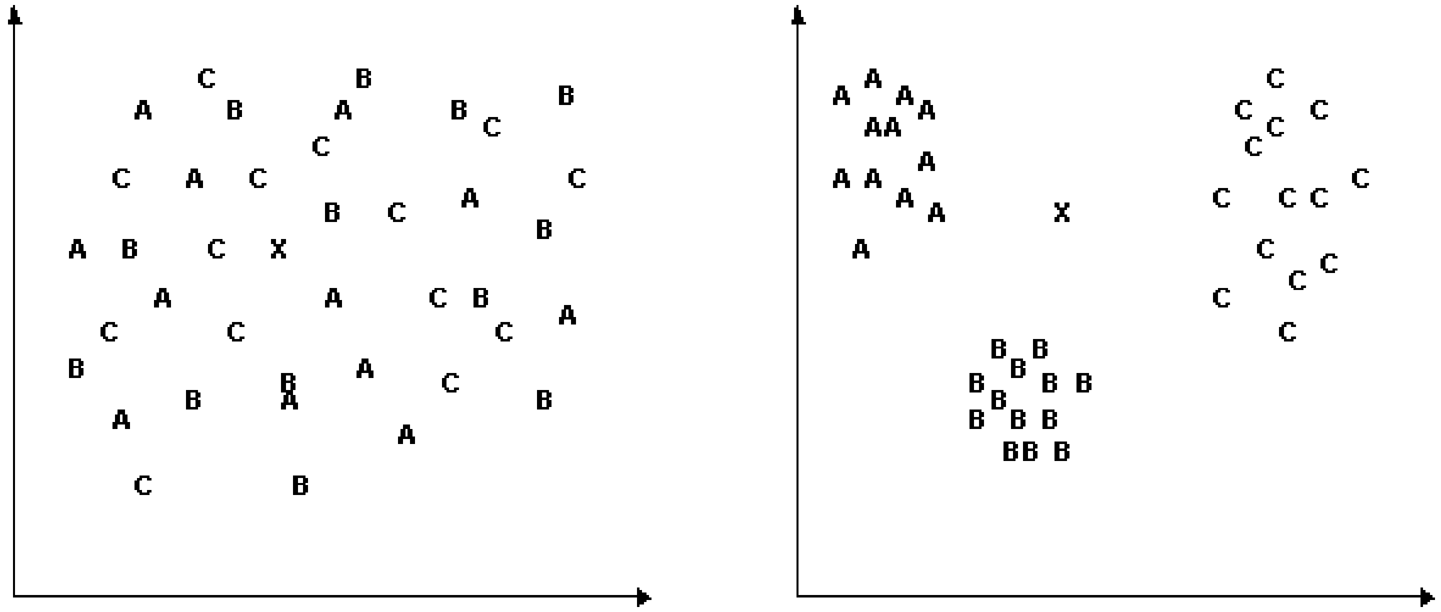
The purpose of digital authorship research is to determine who is the author of an unsigned text in a situation where there are several contenders.

Based on a table of features, the similarity of the texts is determined and their distances are reduced to two dimensions.

Imagine the texts of 3 authors (A, B, C), who may have written text X.

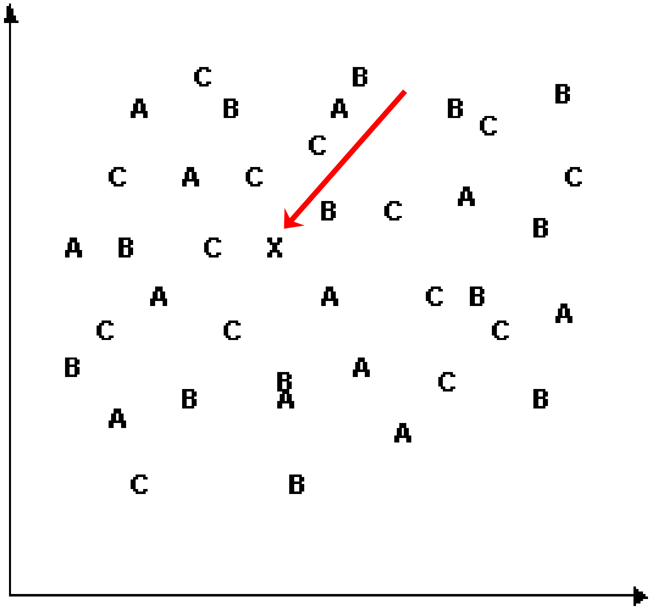
Authorship studies and digital taxonomies

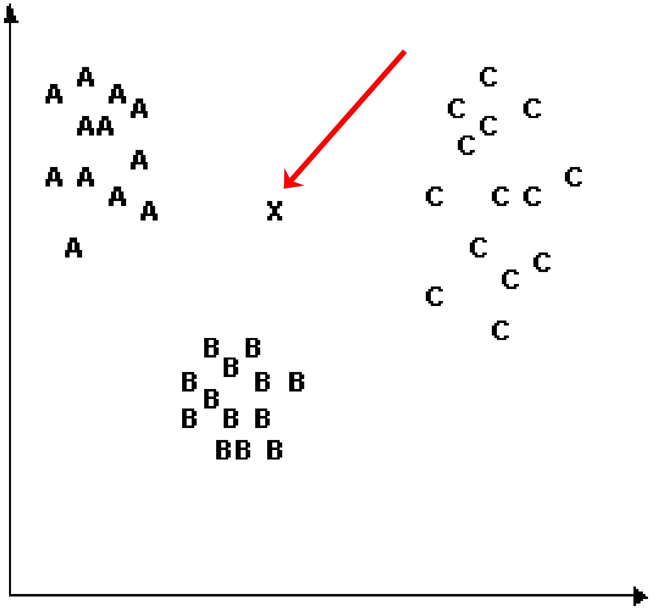
Relationships of texts A, B, C and X can look like this:



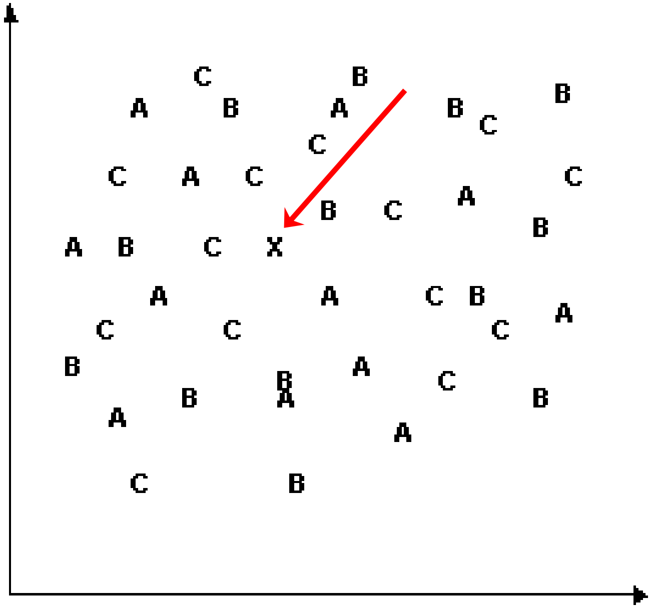
Authorship studies and digital taxonomies

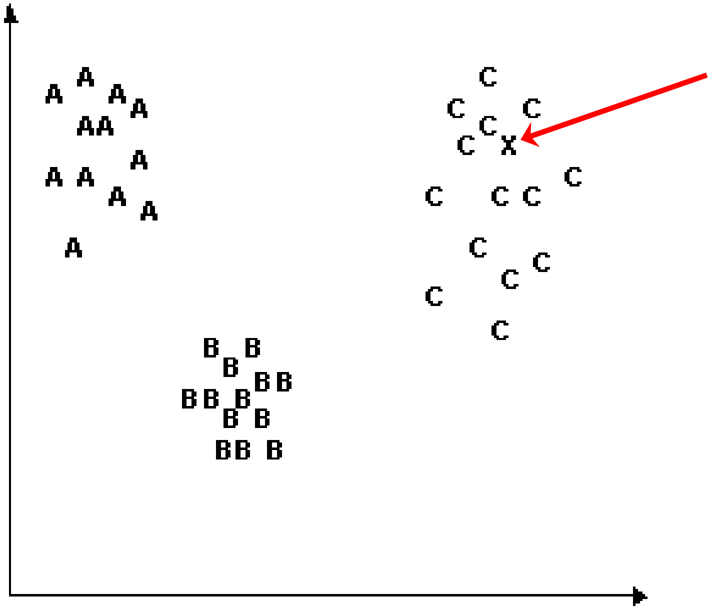
The graph on the left shows a complete lack of similarity between subjects A, B, C, and the place of text X does not allow to suggest its author.
The graph on the right shows the identity of authors A, B, C and the lack of similarity of X to any of them.





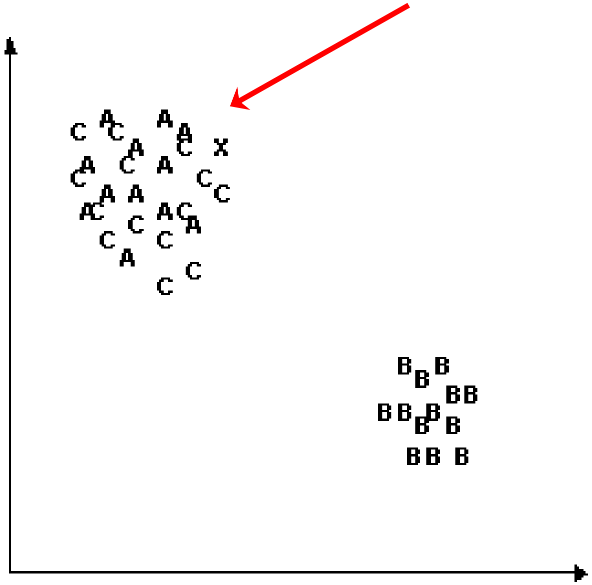
However, the situation may be different.
Here the graph on the right shows the similarity of text X to author C.

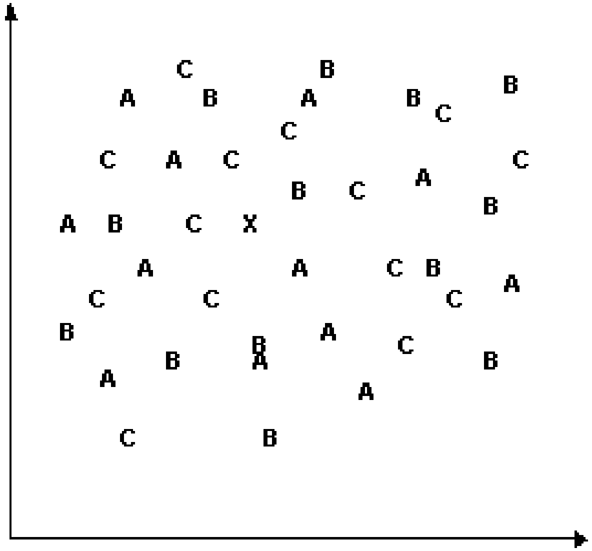




Authorship studies and digital taxonomies

However, the situation may be different.
In this case, the graph on the left shows that probably the authors A. and C and X are probably the same person. And the graph on the right shows that the selection of texts is random.





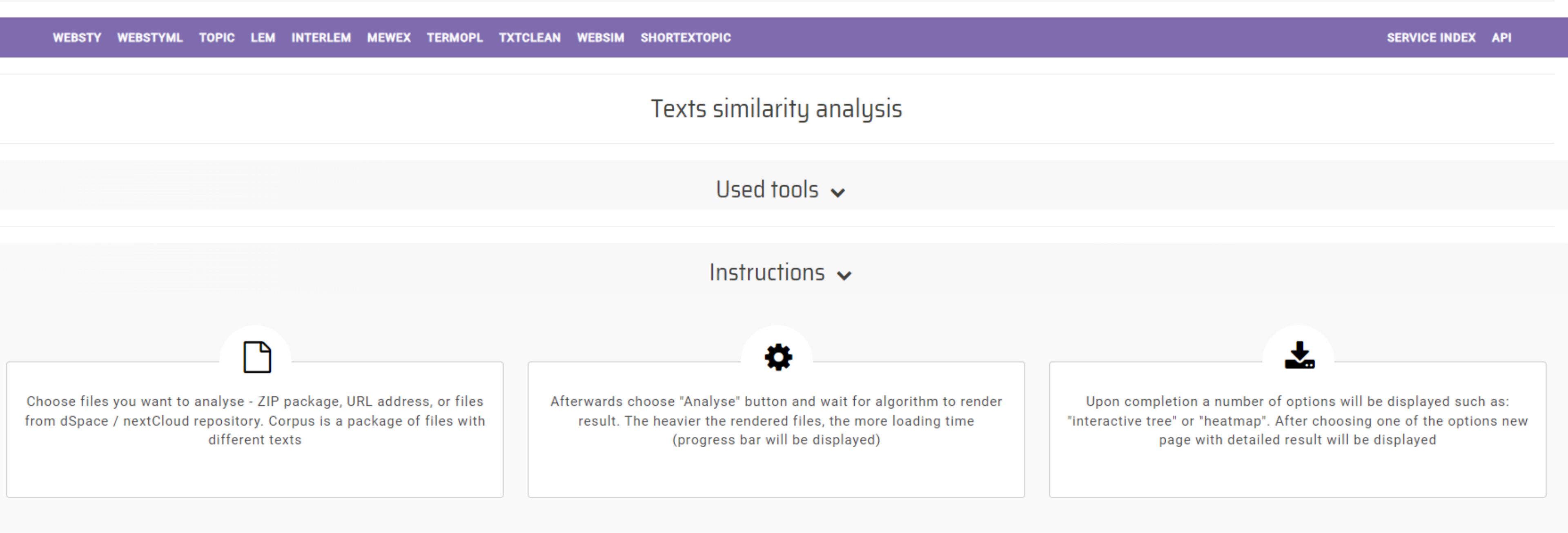
Authorship studies and digital taxonomies

Analysis of real texts can be carried out today with online tools. Such an offer is provided by the CLARIN consortium.

CLARIN

Common Language Resources and Technology Infrastructure

This screenshot shows open and free tools for text analysis.



Texts similarity analysis

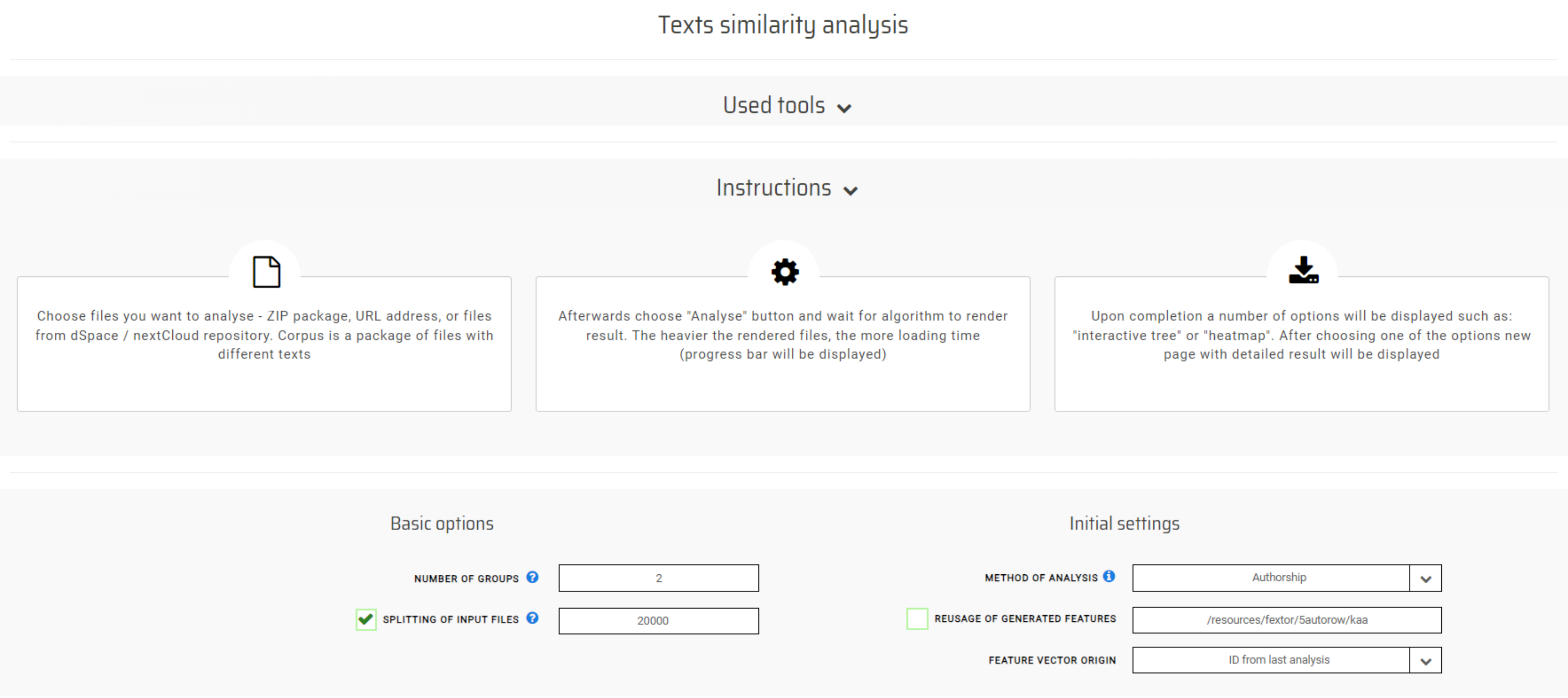
Choose files you want to analyse - ZIP package, URL address, or files from dSpace / nextCloud repository. Corpus is a package of files with different texts

Afterwards choose "Analyse" button and wait for algorithm to render result. The heavier the rendered files, the more loading time (progress bar will be displayed)

Upon completion a number of options will be displayed such as: "interactive tree" or "heatmap". After choosing one of the options new page with detailed result will be displayed

Authorship studies and digital taxonomies

The taxonomy can be conducted with the WebSty module. This module uses multivariate modeling methods.



Texts similarity analysis

Used tools ⌄

Instructions ⌄

Choose files you want to analyse - ZIP package, URL address, or files from dSpace / nextCloud repository. Corpus is a package of files with different texts

Afterwards choose "Analyse" button and wait for algorithm to render result. The heavier the rendered files, the more loading time (progress bar will be displayed)

Upon completion a number of options will be displayed such as: "interactive tree" or "heatmap". After choosing one of the options new page with detailed result will be displayed

Basic options

NUMBER OF GROUPS ? 2

☑ SPLITTING OF INPUT FILES ? 20000

Initial settings

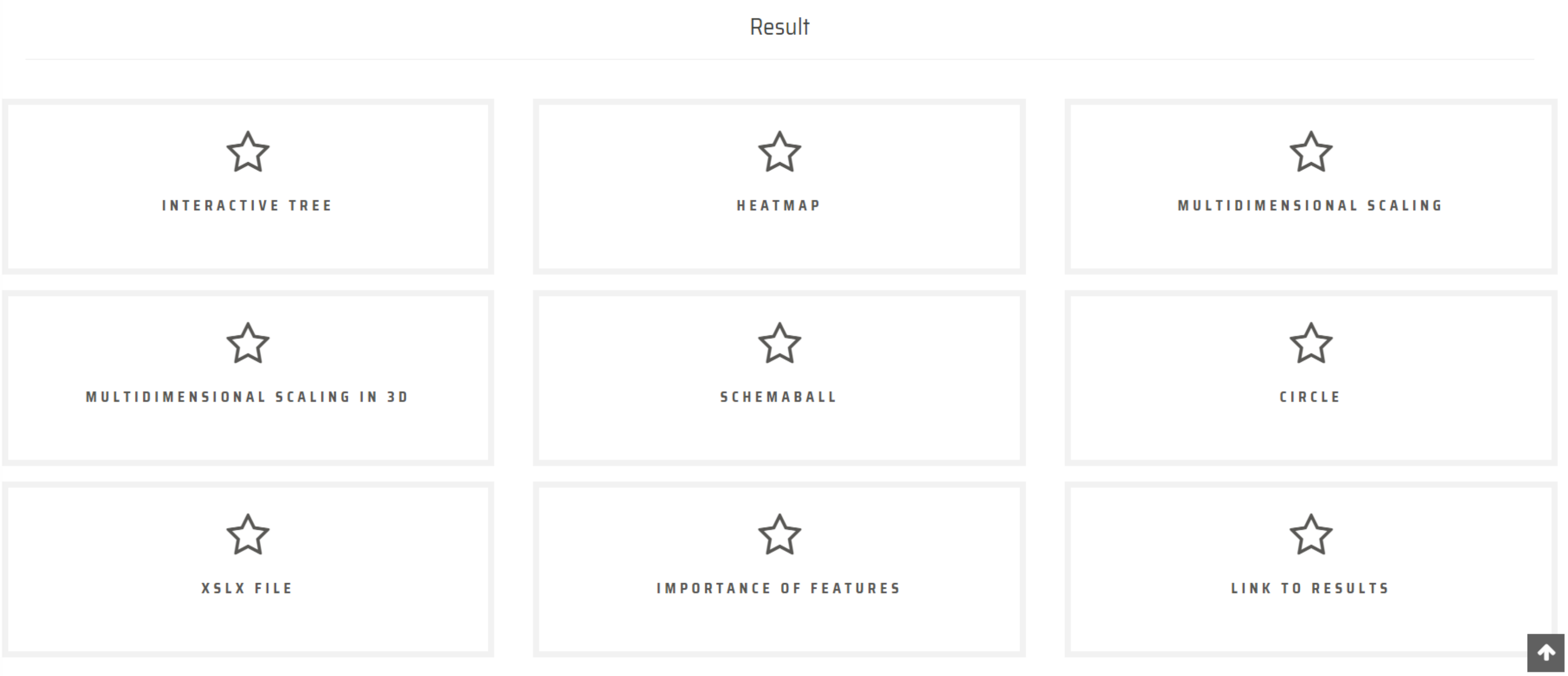
METHOD OF ANALYSIS ℹ Authorship ⌄

☐ REUSAGE OF GENERATED FEATURES /resources/fextor/5autorow/kaa

FEATURE VECTOR ORIGIN ID from last analysis ⌄

Authorship studies and digital taxonomies

The strength of the WebSty system is its rich infographics. This means that the results can be presented in many forms.



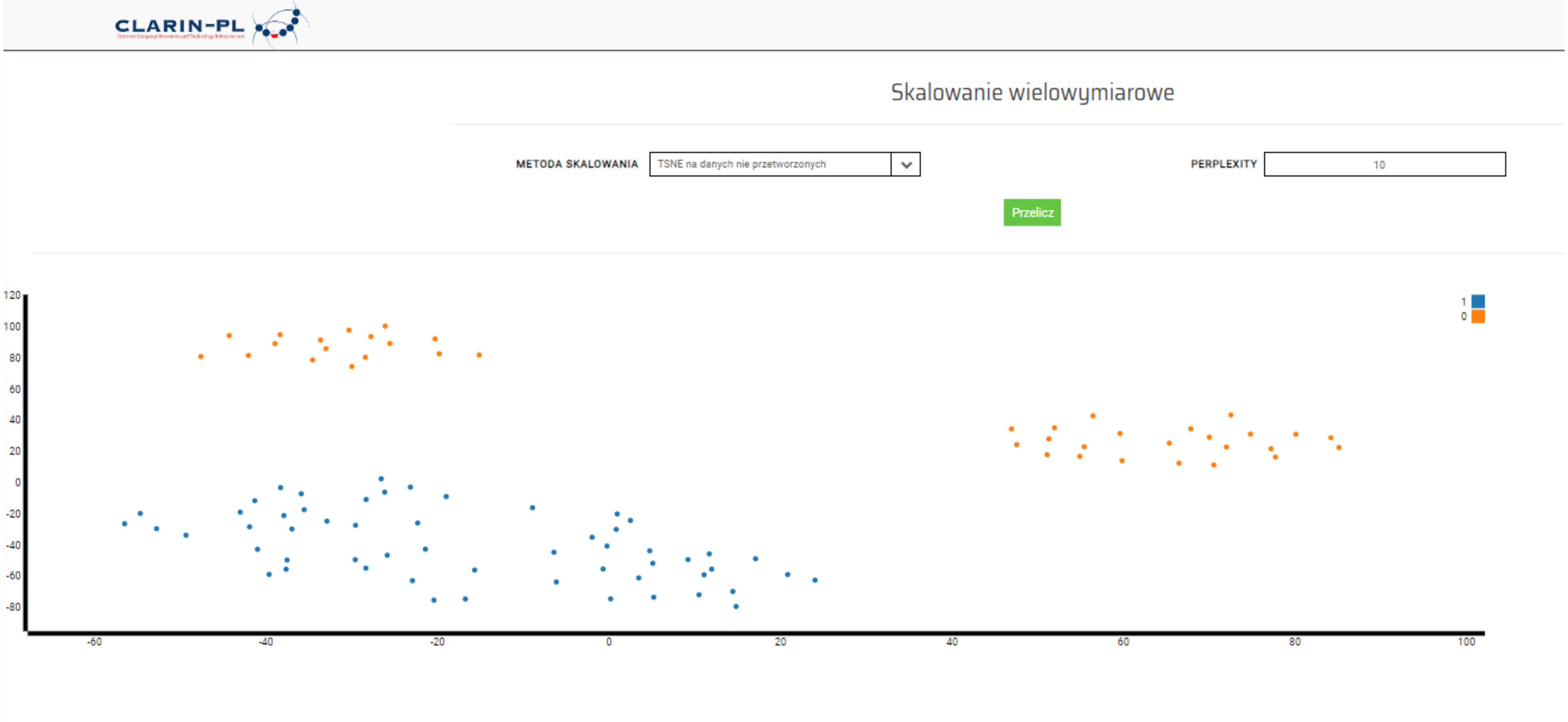
The initial phase of the study of literary texts is to recognize and mark with metatags the parts of speech.

The table on the right shows the statistics of the parts of speech recognized in the sample text (NKJP notation).

tag	all	zeromski	reymont	prus	sienkiewicz	orzeszkowa
interp	395648	53880	97699	57811	110184	76074
ign	133075	34689	5546	2269	10031	80540
qub	119562	15230	27940	14679	39922	21791
conj	101390	12151	23885	11143	33300	20911
adv:pos	40938	5703	11048	3607	11892	8688
adv	38894	5866	7635	4393	12318	8682
subst:sg:nom:m1	34937	4713	7942	5539	12303	4440
praet:sg:m1:perf	34801	4750	8338	4944	12000	4769
subst:sg:gen:f	30894	5210	6221	4140	8581	6742
comp	30601	2993	6894	3759	12719	4236
subst:sg:nom:f	29148	3874	6372	4282	8702	5918
praet:sg:m1:imperf	28842	5070	8000	3105	8544	4123
prep:acc	26873	3722	6536	3572	8545	4498
prep:gen	26053	3969	5288	3193	8262	5341
prep:loc:nwok	25617	4824	5447	2480	7646	5220
subst:sg:acc:f	25181	3637	5654	3516	7606	4768
fin:sg:ter:imperf	24484	2953	4796	4261	8848	3626
prep:loc	23782	3566	5433	2718	7377	4688
prep:gen:nwok	20350	3539	4183	2231	6054	4343
prep:inst:nwok	19655	2601	4842	1949	5492	4771
subst:sg:gen:m3	19524	3802	3803	2505	5112	4302
subst:sg:acc:m3	18526	3099	4621	2580	5049	3177
subst:sg:acc:n	17068	2598	3934	2143	5260	3133
praet:sg:f:perf	16886	1830	3913	2475	3960	4708
inf:imperf	16851	2031	3474	1898	6157	3291
praet:sg:f:imperf	16786	2382	4604	1600	3503	4697
adj:sg:nom:f:pos	15806	2343	3465	1786	4479	3733
subst:sg:nom:n	14921	2435	3002	1643	4692	3149
subst:sg:gen:n	14837	2477	3221	1472	4311	3356
adj:sg:nom:m1:pos	14423	2069	3084	2057	4950	2263
subst:sg:loc:f	13787	2404	2901	1596	4047	2839
inf:perf	13603	1761	2802	1521	5471	2048

Authorship studies and digital taxonomies

The strength of the WebSty system is its rich infographics. This means that results can be presented in multiple forms. Below is an example of multidimensional scaling to the form of a dot plot.



CLARIN-PL

Skalowanie wielowymiarowe

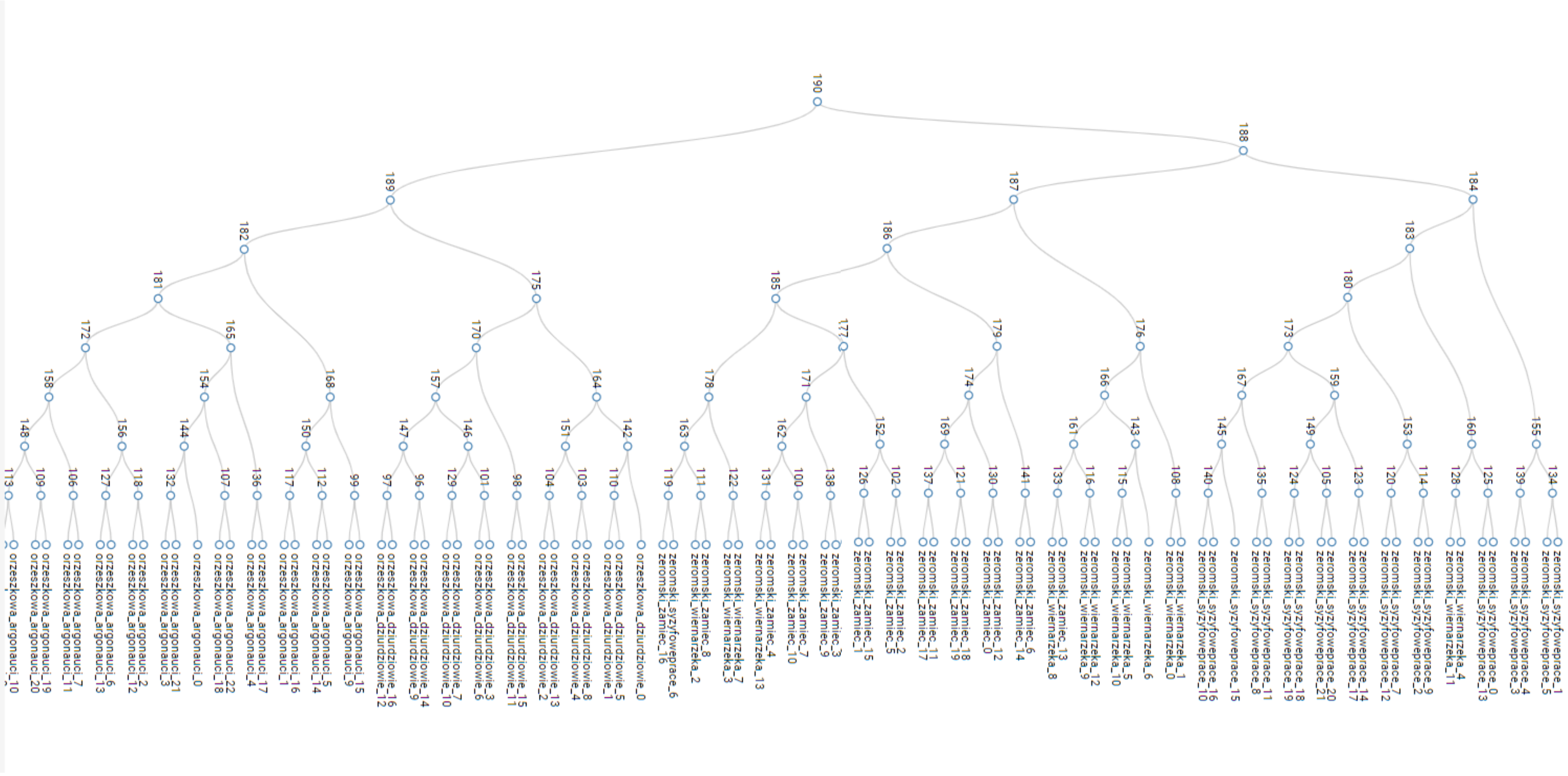
METODA SKALOWANIA: TSNE na danych nie przetworzonych

PERPLEXITY: 10

Przelicz

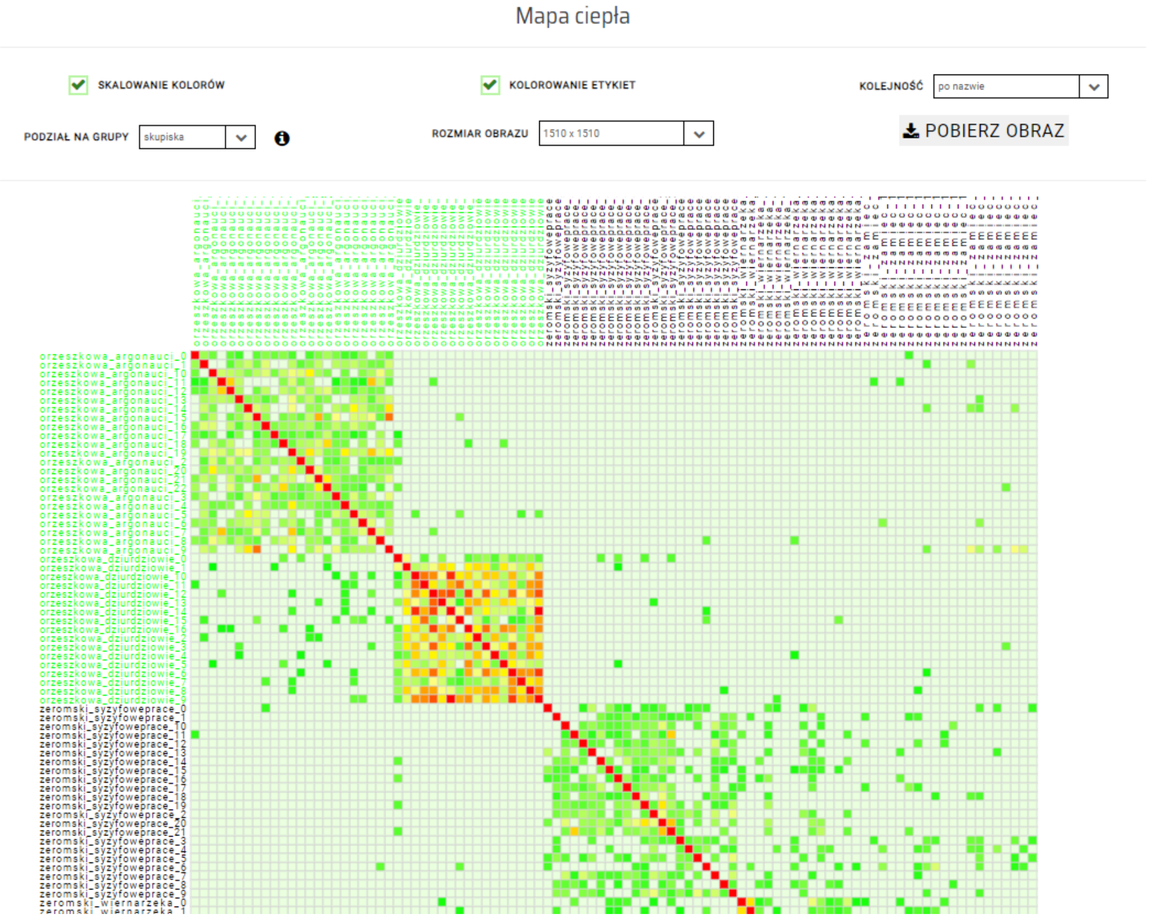
Authorship studies and digital taxonomies

Automatic analysis of the texts leads to the creation of a dendrogram. Here you can see the dendrogram of 5 authors and dozens of novels.



Authorship studies and digital taxonomies

Relationships captured as a dendrogram or scatterplot can also be shown in the form of a heat map. A heat map is a matrix of correlations that resembles a thermal image. The darker the color, the greater the correlation.



Mapa ciepła

☑ SKALOWANIE KOLORÓW

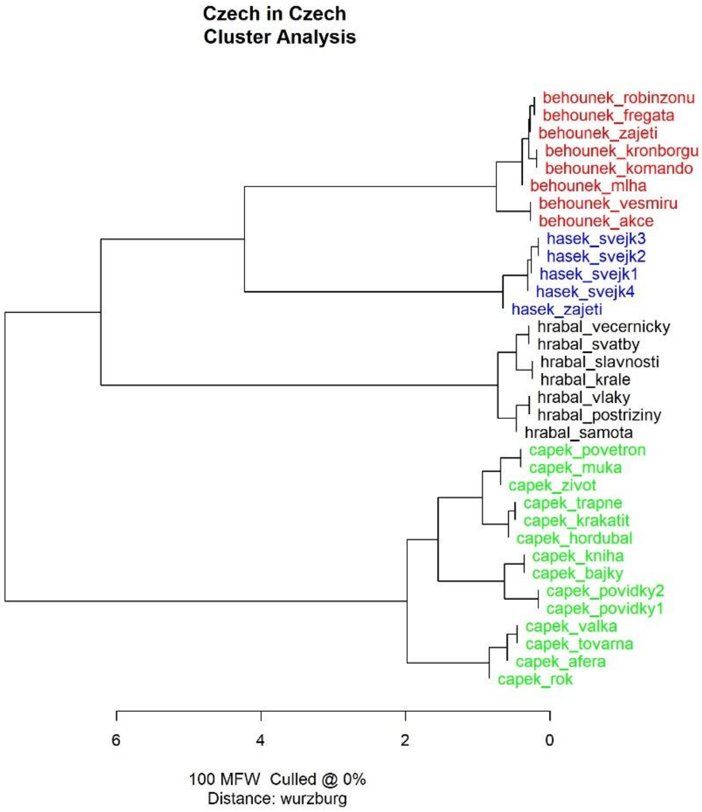
☑ KOLOROWANIE ETYKIET

KOLEJNOŚĆ: po nazwie

PODZIAŁ NA GRUPY: skupiska

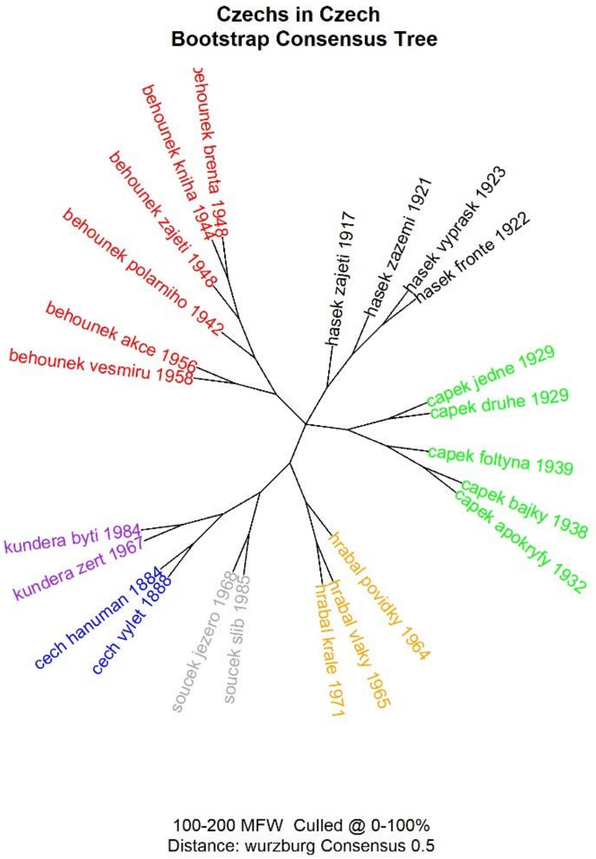
ROZMIAR OBRAZU: 1510 x 1510

The graphic below shows the text relationships of novels by Czech authors.



Authorship studies and digital taxonomies

The infographic shows the text relationships of novels by Czech authors in the form of a radar chart.



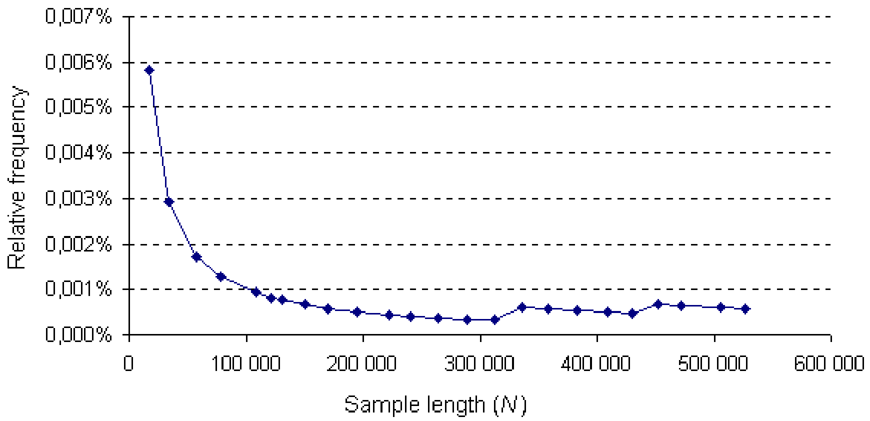
Authorship studies and digital taxonomies

Vocabulary analysis to discover authorship can be risky if the samples studied are of different lengths.

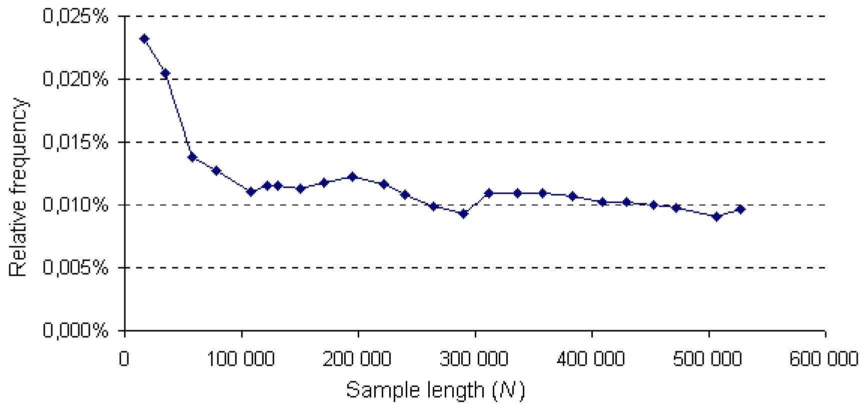
The graphs show how the relative frequency of lexemes ‘luty’ (February) and ‘praca’ (work) changes as the length of the text increases.

This means that texts with significantly different lengths should not be compared.

leksem 'luty'

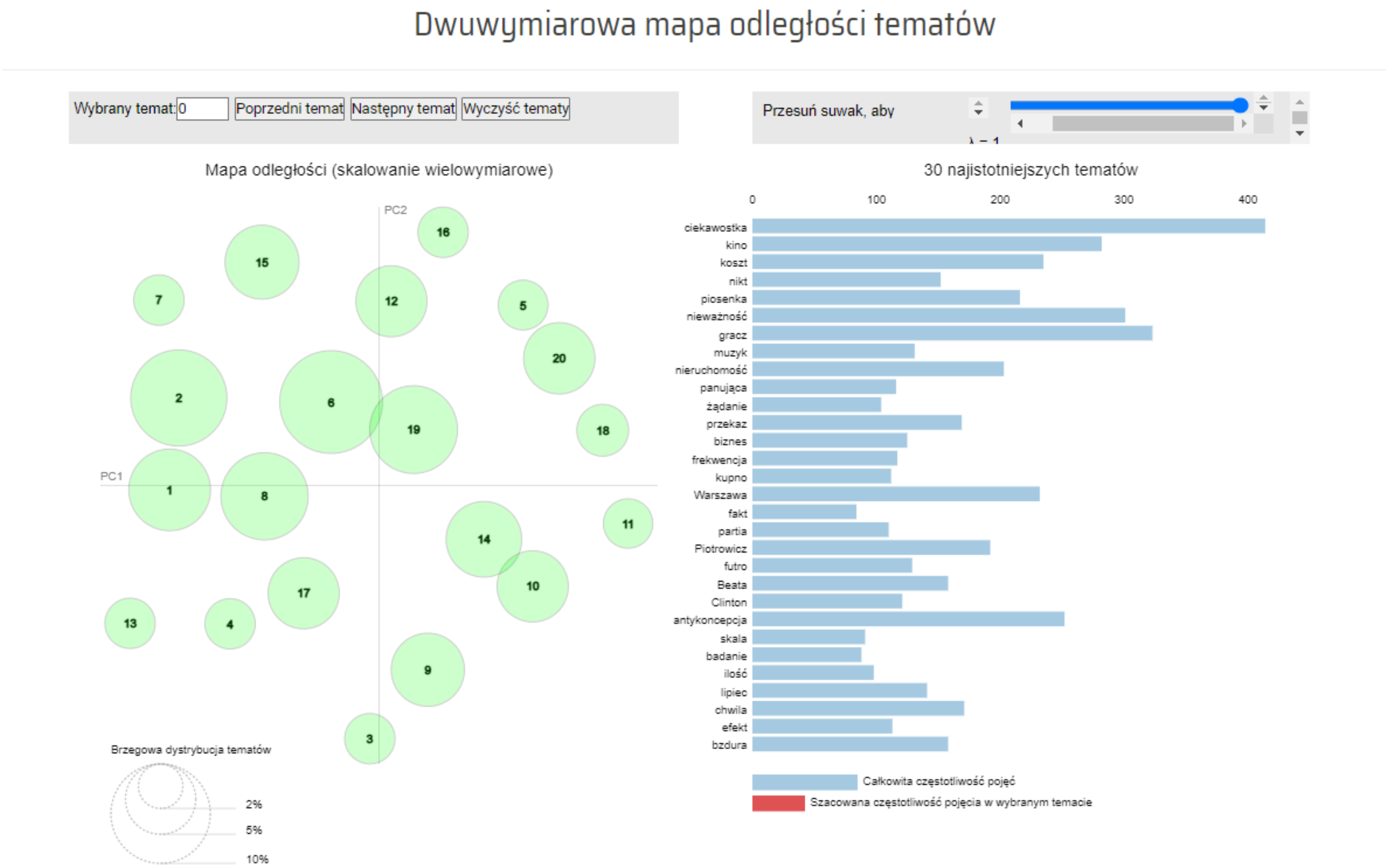


leksem 'praca'



Thematic analysis of texts

It is also possible to build thematic maps of texts, containing topics, or clusters of topic words, semantically and distributively related to each other.



Dwuwymiarowa mapa odległości tematów

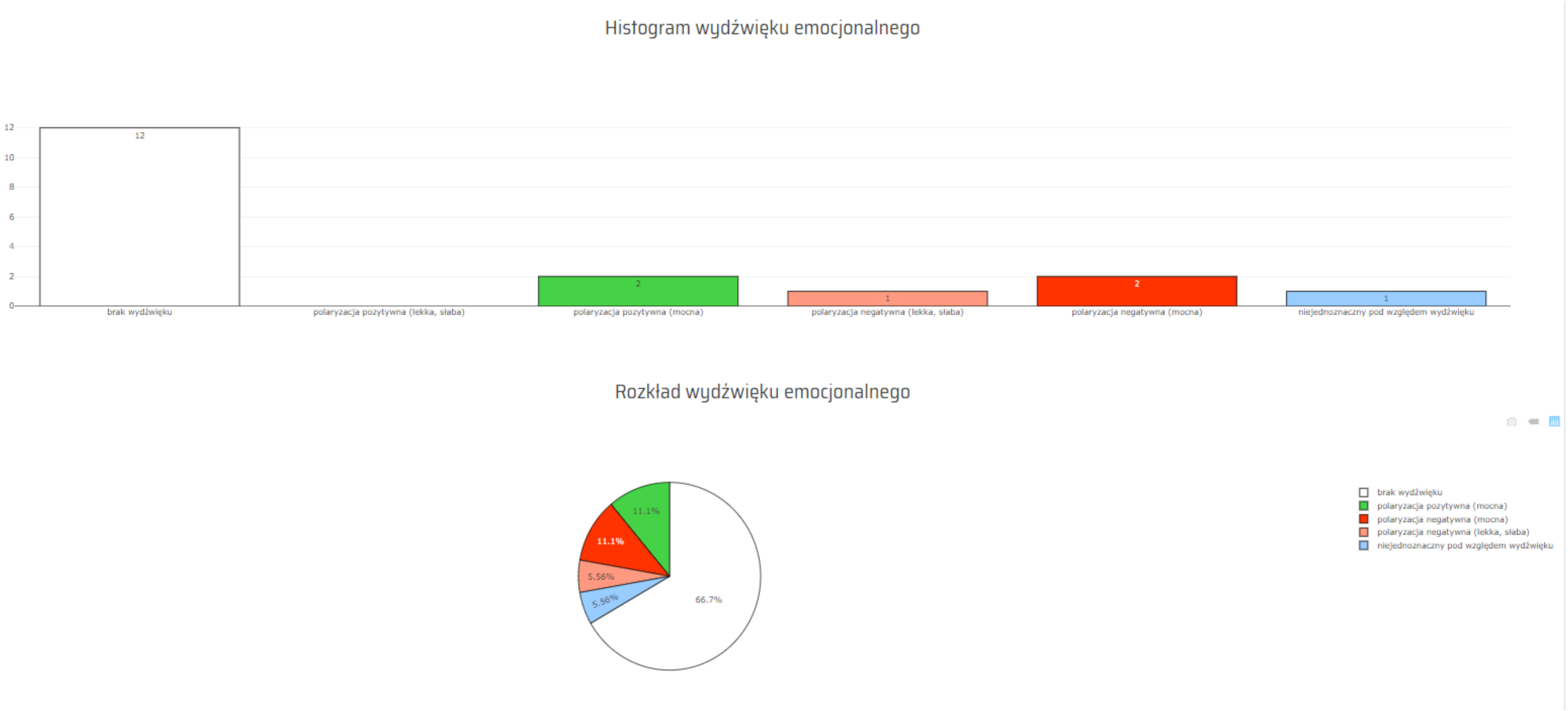
Wybrany temat: 0

Analysis of sentiment in text

The emotional resonance of a text is a simplified "emotional picture" of the text, based on vocabulary analysis. The assignment of emotional value to individual words is done automatically based on bases created from studies of human reactions to text.

Analysis of sentiment in text

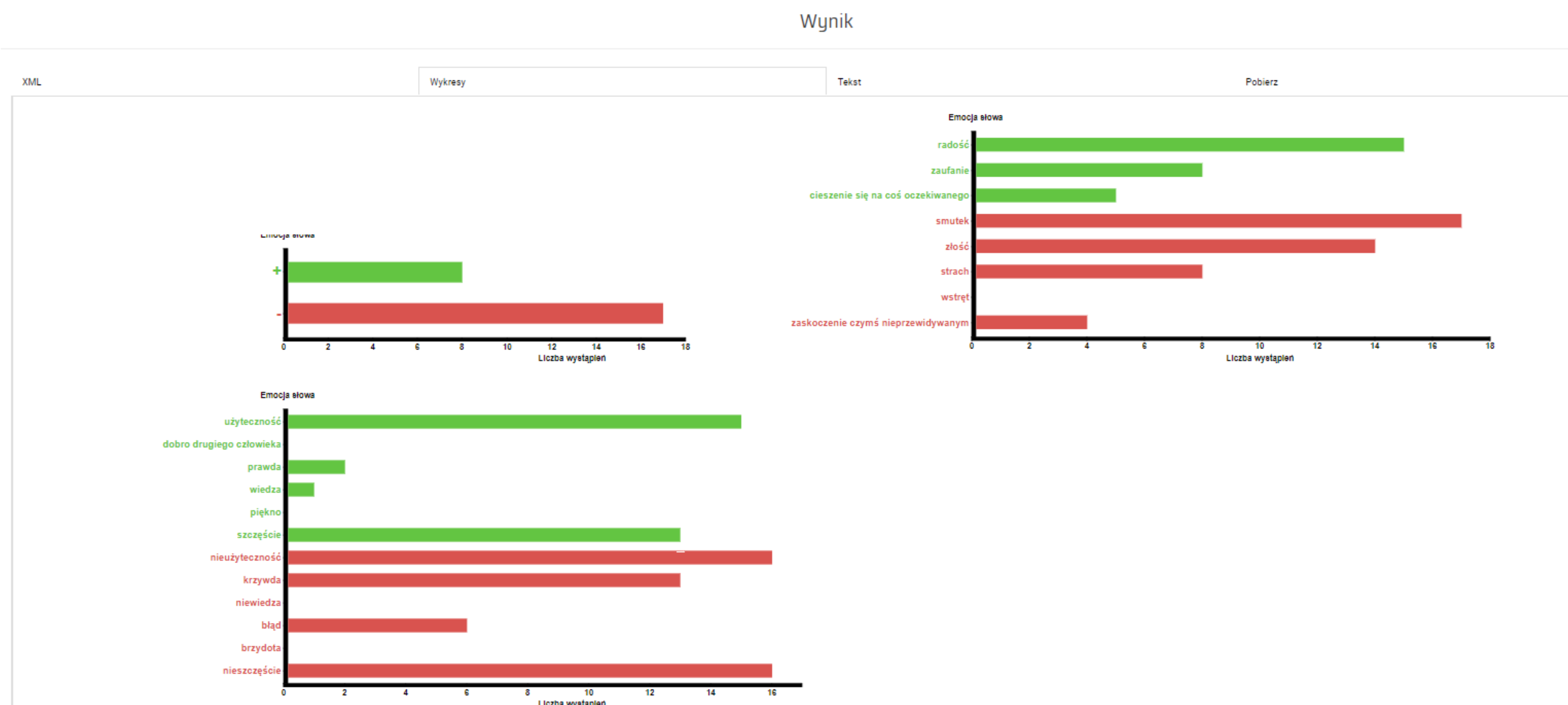
The infographic below shows the distribution of emotive expressions.



Uniwersytet Wrocławski

Analysis of sentiment in text

Rozkład wyrażeń emotywnych może być mniej lub bardziej szczegółowy. Główne kategorie to wydźwięk pozytywny / neutralny / negatywny. Oprócz tego rozpoznawane są szczegółowe rodzaje emocji.



Digital edition

Digital editing of source texts involves the construction of resources that contain most of the source texts along with annotation and accounts.

Especially suitable for this are dated collections that can be placed on a timeline.

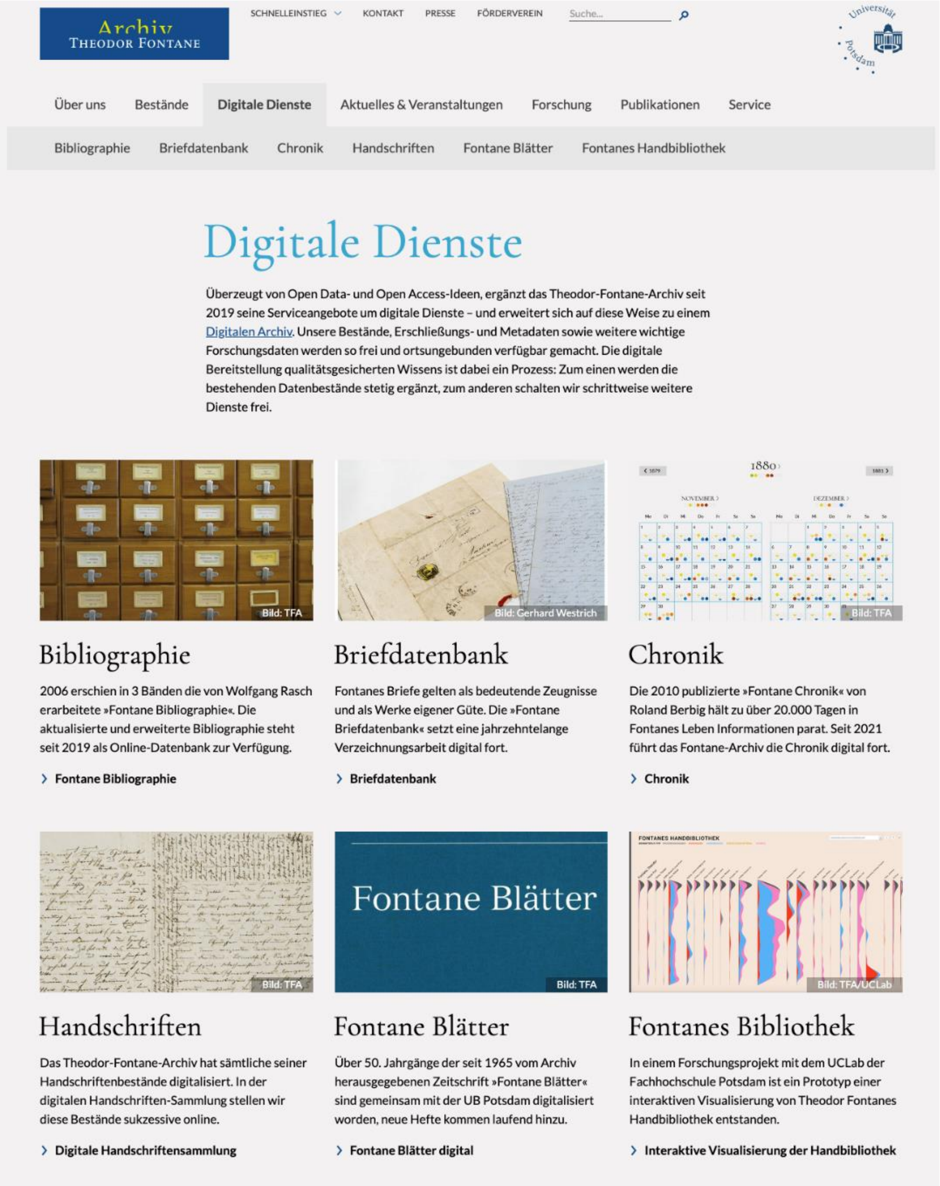
Corpora of letters are ideally suited for this purpose.

Digital edition

Digital literary studies operates on infrastructures.

Such an infrastructure can be a thematic institution that dedicates its efforts to the analysis of a particular researcher.

The infographic shows the offerings of the Theodor Fontane Institute.



Digitale Dienste

Überzeugt von Open Data- und Open Access-Ideen, ergänzt das Theodor-Fontane-Archiv seit 2019 seine Serviceangebote um digitale Dienste – und erweitert sich auf diese Weise zu einem Digitalen Archiv. Unsere Bestände, Erschließungs- und Metadaten sowie weitere wichtige Forschungsdaten werden so frei und ortsungebunden verfügbar gemacht. Die digitale Bereitstellung qualitätsgesicherten Wissens ist dabei ein Prozess: Zum einen werden die bestehenden Datenbestände stetig ergänzt, zum anderen schalten wir schrittweise weitere Dienste frei.

Bibliographie

2006 erschien in 3 Bänden die von Wolfgang Rasch erarbeitete »Fontane Bibliographie«. Die aktualisierte und erweiterte Bibliographie steht seit 2019 als Online-Datenbank zur Verfügung.

› Fontane Bibliographie

Briefdatenbank

Fontanes Briefe gelten als bedeutende Zeugnisse und als Werke eigener Güte. Die »Fontane Briefdatenbank« setzt eine jahrzehntelange Verzeichnungsarbeit digital fort.

› Briefdatenbank

Chronik

Die 2010 publizierte »Fontane Chronik« von Roland Berbig hält zu über 20.000 Tagen in Fontanes Leben Informationen parat. Seit 2021 führt das Fontane-Archiv die Chronik digital fort.

› Chronik

Handschriften

Das Theodor-Fontane-Archiv hat sämtliche seiner Handschriftenbestände digitalisiert. In der digitalen Handschriften-Sammlung stellen wir diese Bestände sukzessive online.

› Digitale Handschriftensammlung

Fontane Blätter

Über 50. Jahrgänge der seit 1965 vom Archiv herausgegebenen Zeitschrift »Fontane Blätter« sind gemeinsam mit der UB Potsdam digitalisiert worden, neue Hefte kommen laufend hinzu.

› Fontane Blätter digital

Fontanes Bibliothek

In einem Forschungsprojekt mit dem UCLab der Fachhochschule Potsdam ist ein Prototyp einer interaktiven Visualisierung von Theodor Fontanes Handbibliothek entstanden.

› Interaktive Visualisierung der Handbibliothek

Digital edition

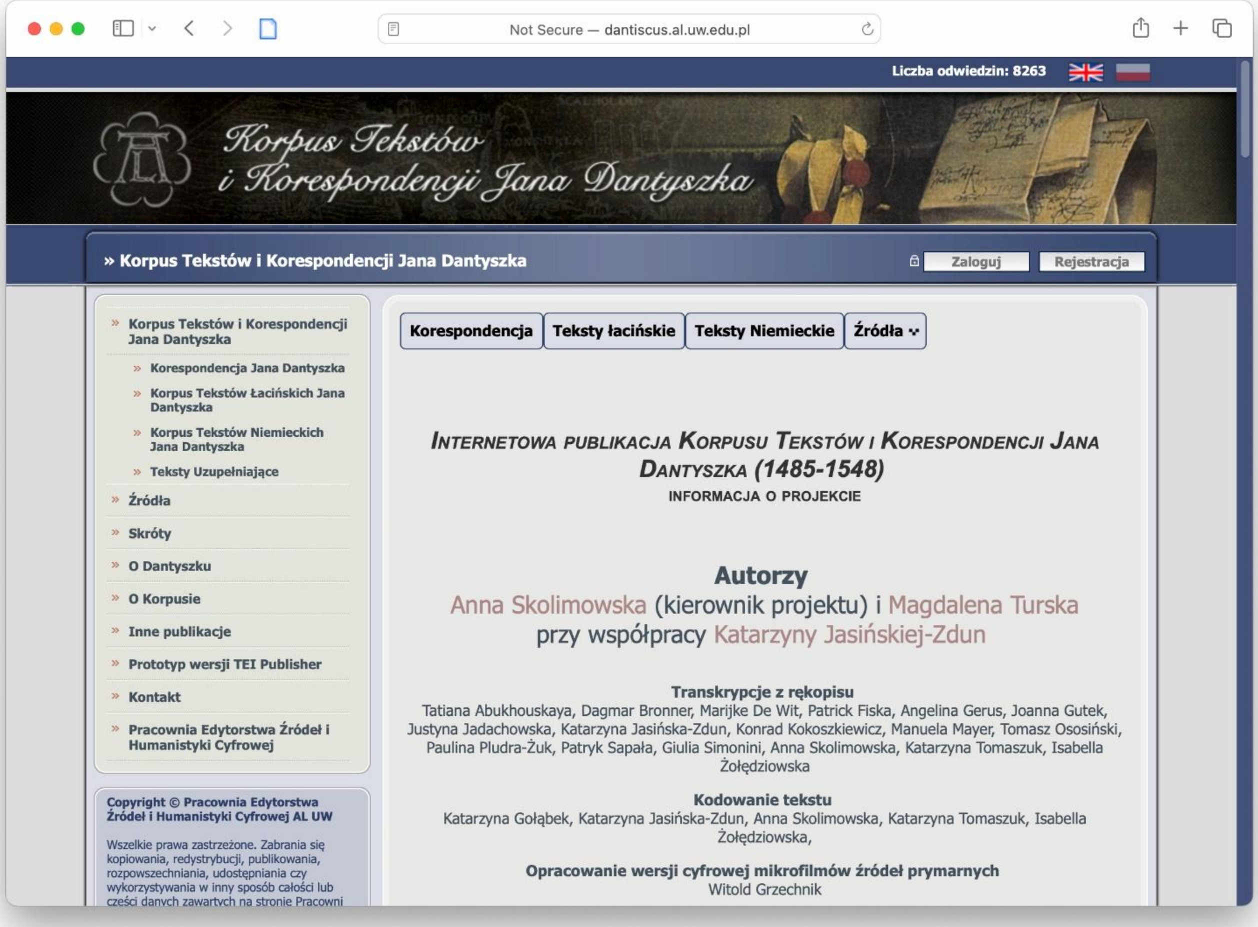
Digital editing of source texts involves the construction of resources that contain most of the source texts along with annotation and accounts.

Especially suitable for this are dated collections that can be placed on a timeline.

Corpora of letters are ideally suited for this purpose.

Digital edition

The graphic shows a portal with Jan Dantyszek's correspondence. Today it is the best portal of its kind in Poland.


Not Secure — dantiscus.al.uw.edu.pl
Liczba odwiedzin: 8263
Korpus Tekstów i Korespondencji Jana Dantyszka
» Korpus Tekstów i Korespondencji Jana Dantyszka
Zaloguj
Rejestracja
» Korpus Tekstów i Korespondencji Jana Dantyszka
» Korespondencja Jana Dantyszka
» Korpus Tekstów Łacińskich Jana Dantyszka
» Korpus Tekstów Niemieckich Jana Dantyszka
» Teksty Uzupełniające
» Źródła
» Skróty
» O Dantyszku
» O Korpusie
» Inne publikacje
» Prototyp wersji TEI Publisher
» Kontakt
» Pracownia Edytorstwa Źródeł i Humanistyki Cyfrowej
Korespondencja
Teksty łacińskie
Teksty Niemieckie
Źródła

Copyright © Pracownia Edytorstwa Źródeł i Humanistyki Cyfrowej AL UW

Wszelkie prawa zastrzeżone. Zabrania się kopiowania, redystrybucji, publikowania, rozpowszechniania, udostępniania czy wykorzystywania w inny sposób całości lub części danych zawartych na stronie Pracowni

INTERNETOWA PUBLIKACJA KORPUSU TEKSTÓW I KORESPONDENCJI JANA DANTYSZKA (1485-1548)

INFORMACJA O PROJEKCIE

Autorzy

Anna Skolimowska (kierownik projektu) i Magdalena Turska przy współpracy Katarzyny Jasińskiej-Zdun

Transkrypcje z rękopisu

Tatiana Abukhouskaya, Dagmar Bronner, Marijke De Wit, Patrick Fiska, Angelina Gerus, Joanna Gutek, Justyna Jadachowska, Katarzyna Jasińska-Zdun, Konrad Kokoszkiewicz, Manuela Mayer, Tomasz Ososiński, Paulina Pludra-Żuk, Patryk Sapała, Giulia Simonini, Anna Skolimowska, Katarzyna Tomaszuk, Isabella Żołędziowska

Kodowanie tekstu

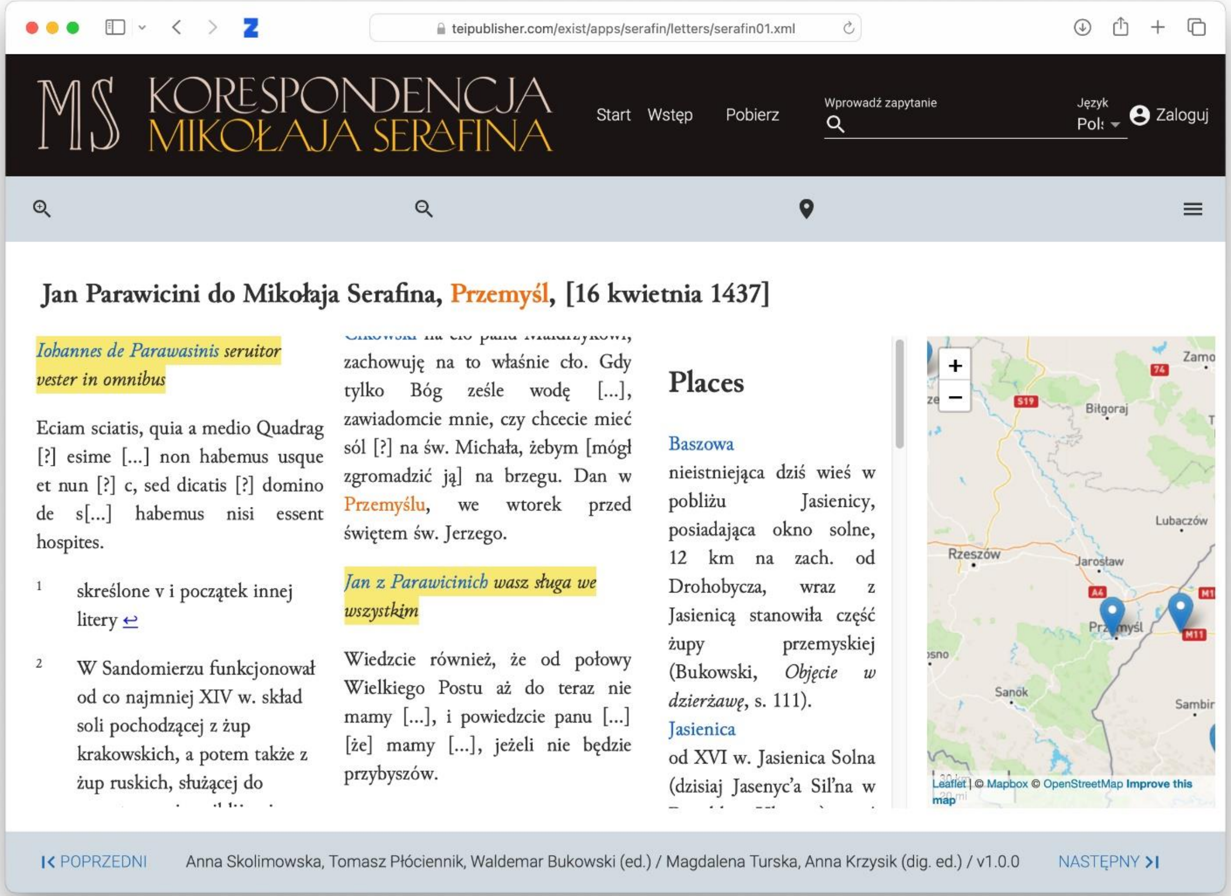
Katarzyna Gołąbek, Katarzyna Jasińska-Zdun, Anna Skolimowska, Katarzyna Tomaszuk, Isabella Żołędziowska,

Opracowanie wersji cyfrowej mikrofilmów źródeł prymarnych

Witold Grzechnik

Digital edition

The graphic shows a portal with the correspondence of Nicholas Serafin. Texts include annotation of author's corrections, and are linked to the map.



teipublisher.com/exist/apps/serafin/letters/serafin01.xml
MS KORESPONDENCJA MIKOŁAJA SERAFINA
Start
Wstęp
Pobierz
Wprowadź zapytanie
Język
Pol:
Zaloguj
+
−
Zamo
74
S19
Biłgoraj
Lubaczów
Rzeszów
Jarosław
A4
M1
Przemyśl
M11
Sanok
Sambir
Leaflet | © Mapbox © OpenStreetMap Improve this map

Jan Parawicini do Mikołaja Serafina, Przemyśl, [16 kwietnia 1437]

Iohannes de Parawasinis seruitor vester in omnibus

Eciam sciatis, quia a medio Quadrag [?] esime [...] non habemus usque et nun [?] c, sed dicatis [?] domino de s[...] habemus nisi essent hospites.

zachowuję na to właśnie cło. Gdy tylko Bóg ześle wodę [...], zawiadomcie mnie, czy chcecie mieć sól [?] na św. Michała, żebym [mógł zgromadzić ją] na brzegu. Dan w Przemyślu, we wtorek przed świętem św. Jerzego.

Jan z Parawicinich wasz sługa we wszystkim

Wiedzcie również, że od połowy Wielkiego Postu aż do teraz nie mamy [...], i powiedzcie panu [...] [że] mamy [...], jeżeli nie będzie przybyszów.

[1] skreślone v i początek innej litery ↩

[2] W Sandomierzu funkcjonował od co najmniej XIV w. skład soli pochodzącej z żup krakowskich, a potem także z żup ruskich, służącej do

Places

Baszowa
nieistniejąca dziś wieś w pobliżu Jasienicy, posiadająca okno solne, 12 km na zach. od Drohobycza, wraz z Jasienicą stanowiła część żupy przemyskiej (Bukowski, *Objęcie w dzierżawę*, s. 111).

Jasienica
od XVI w. Jasienica Solna (dzisiaj Jasenyc'a Sil'na w

|< POPRZEDNI Anna Skolimowska, Tomasz Płóciennik, Waldemar Bukowski (ed.) / Magdalena Turska, Anna Krzysik (dig. ed.) / v1.0.0 NASTĘPNY >|