Biuletyn
Polonistyczny

20.06.2024

Tekst, korpus, wykres. Cyfrowa analiza literatury – przydatne narzędzia

---

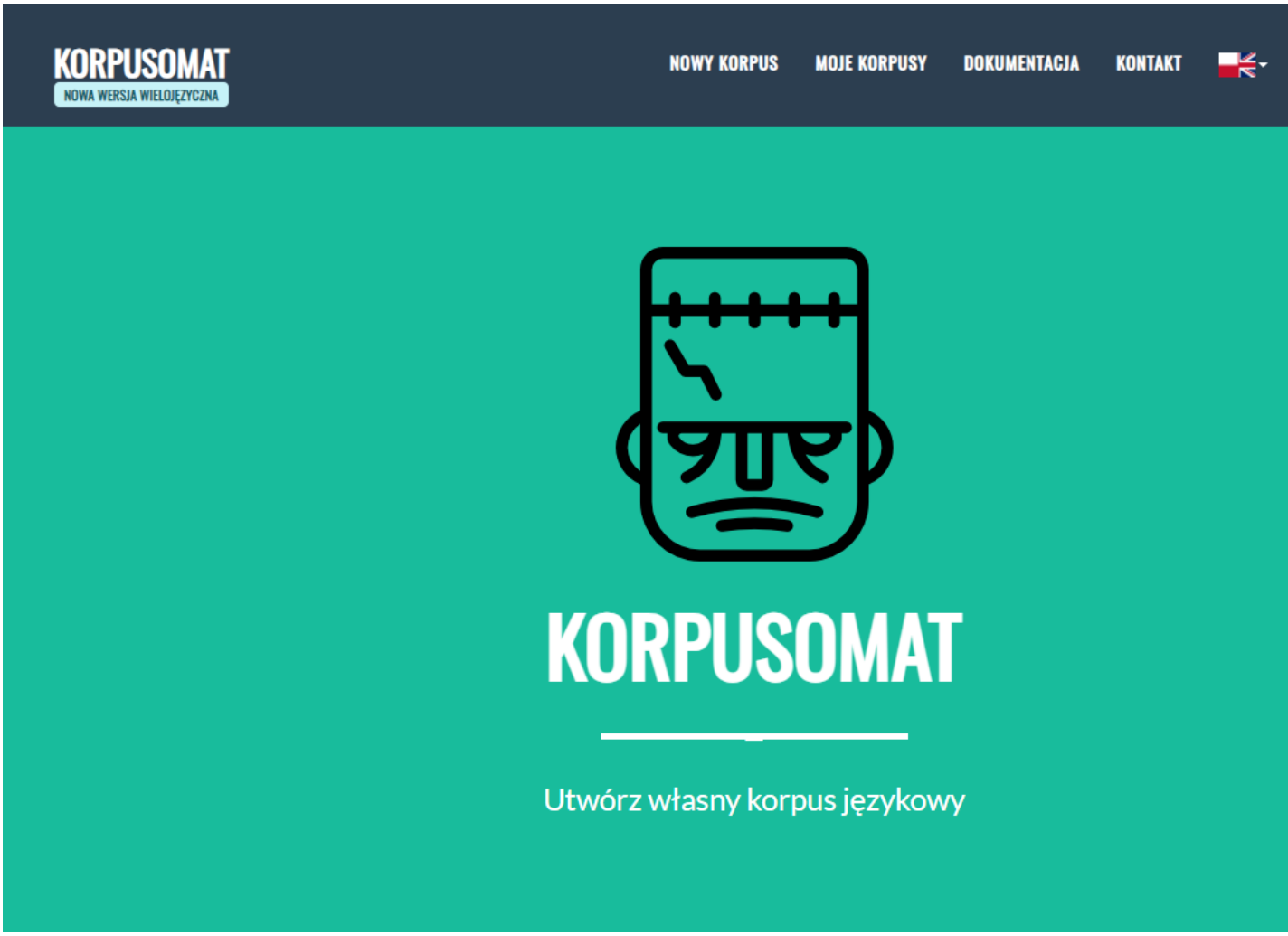
Czy (a może jak?) komputer może czytać literaturę? Choć poziom „rozumienia” tekstu literackiego przez komputer nie odpowiada temu, jak rozumie go człowiek, to narzędzia cyfrowe mogą wydatnie wspomóc różnego rodzaju badania literackie. Dziś nie musimy już ręcznie liczyć wystąpień interesującego nas słowa w tekście. Dzięki zaawansowanym wyszukiwarkom możemy wyszukiwać całe frazy i podobnie zbudowane fragmenty, a lista słów użytych w badanym przez nas utworze może powstać w kilka chwil.

Wiele zaawansowanych projektów z zakresu cyfrowych badań literackich opiera się o stworzenie własnych, szytych na miarę narzędzi. Jednak dzięki istnieniu infrastruktur badawczych, udostępniających narzędzia za darmo, wiele analiz można przeprowadzić bez konieczności tworzenia specjalnych rozwiązań. Oto kilka przydatnych serwisów dostarczanych przez infrastrukturę CLARIN-PL, które można wykorzystać w swojej codziennej pracy. Żadne z wybranych narzędzi nie wymaga instalacji — możemy używać ich bezpośrednio w przeglądarce. Aby tworzyć i przetwarzać korpusy, musimy tylko założyć darmowe konto i się zalogować.

Na korpusy — Korpusomat!

Podstawą każdego badania, które obejmuje cyfrową analizę tekstów, jest stworzenie korpusu — czyli zbioru tekstów, który poddamy analizie. To kluczowy etap — dobór materiału będzie miał wpływ na wyniki, więc warto zastanowić się, jakie kryteria muszą spełniać teksty, które uwzględnimy w korpusie, ile powinno ich być, by wyniki były miarodajne, oraz czy dobrana próba rzeczywiście będzie w stanie odpowiedzieć nam na stawiane pytania badawcze.

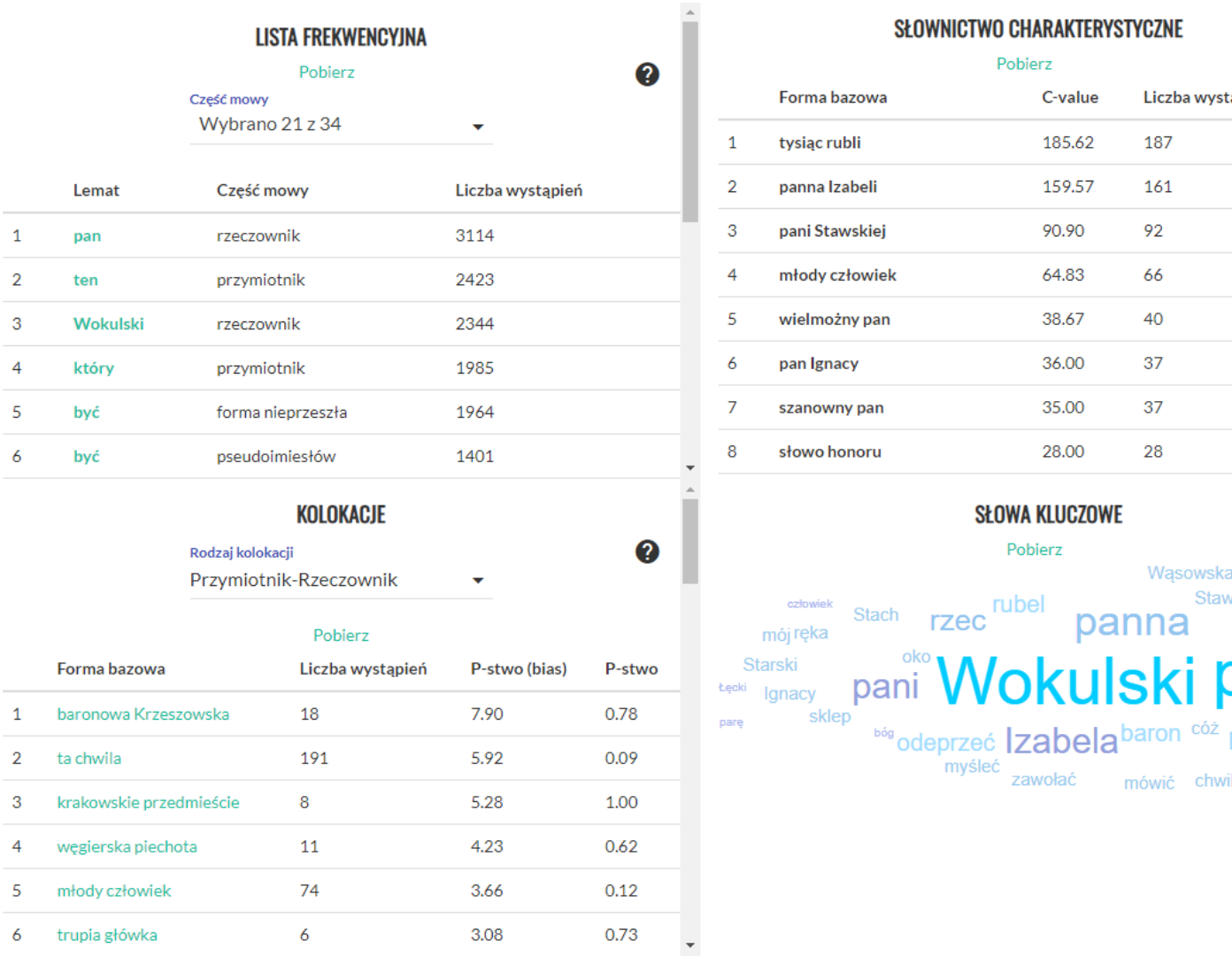
Zebrane teksty najlepiej będzie umieścić w narzędziu, które pozwoli nam na przeszukiwanie korpusu i przynajmniej wstępne analizy. Przedstawiam więc **Korpusomat — narzędzie**, które w bardzo wielu wypadkach będzie wystarczające dla uzyskania podstawowych analiz i zaawansowanego przeszukiwania.



W Korpusomacie każdy może stworzyć własny korpus tekstów, dodając pliki w formacie .txt. Możliwe jest opisanie ich metadanymi, z których część (autor i tytuł) jest obowiązkowa, ale większość pól możemy utworzyć sami. Narzędzie może też samo pobierać dla nas teksty ze wskazanych stron internetowych. Metadane, oprócz dostarczenia informacji o tekstach, posłużą później do filtrowania wyników, dlatego warto dobrze przemyśleć zestaw, którym chcemy się posłużyć (więcej o istotności metadanych i dbałości o ich jakość pisał Cezary Rosiński we wcześniejszym wpisie). Metadane mogą też być odczytywane automatycznie. Oczywiście ich edycja będzie możliwa na każdym etapie pracy z korpusem.

Pliki przechowywane w korpusomacie zostają poddane procesom tagowania i lematyzacji, co oznacza, że każde słowo otrzymuje przypisany tag morfosyntaktyczny (zestaw cech, zawierających m.in. informację o części mowy i wartościach odpowiednich klas gramatycznych) oraz lemat (czyli jednostkę słownikową – np. wyraz „kocie” zostanie przyporządkowany do lematu „kot”). To umożliwia korzystanie z zaawansowanej wyszukiwarki korpusowej, która pozwala na znalezienie w korpusie praktycznie wszystkiego. Określenia koloru niebieskiego, opisujące rzeczownik w rodzaju męskim, w liczbie pojedynczej? Takie i inne, równie złożone zapytania, można zbudować korzystając z konstruktora zapytań. Wyniki można filtrować i grupować, np. według metadanych.

Dodatkowo Korpusomat dysponuje możliwością wyświetlenia statystyk korpusu, gdzie możemy znaleźć np. informację o słownictwie charakterystycznym, listę frekwencyjną, zestawienie kolokacji i słów kluczowych, rozkład słów kluczowych czy podstawowy wykres stylometryczny.



LISTA FREKWENCYJNA

Pobierz

Część mowy

Wybrano 21 z 34

	Lemat	Część mowy	Liczba wystąpień
1	pan	rzeczownik	3114
2	ten	przymiotnik	2423
3	Wokulski	rzeczownik	2344
4	który	przymiotnik	1985
5	być	forma nieprzeszła	1964
6	być	pseudoimiesłów	1401

KOLOKACJE

Rodzaj kolokacji

Przymiotnik-Rzeczownik

Pobierz

	Forma bazowa	Liczba wystąpień	P-stwo (bias)	P-stwo
1	baronowa Krzeszowska	18	7.90	0.78
2	ta chwila	191	5.92	0.09
3	krakowskie przedmieście	8	5.28	1.00
4	węgierska piechota	11	4.23	0.62
5	młody człowiek	74	3.66	0.12
6	trupia główka	6	3.08	0.73

SŁOWNICTWO CHARAKTERYSTYCZNE

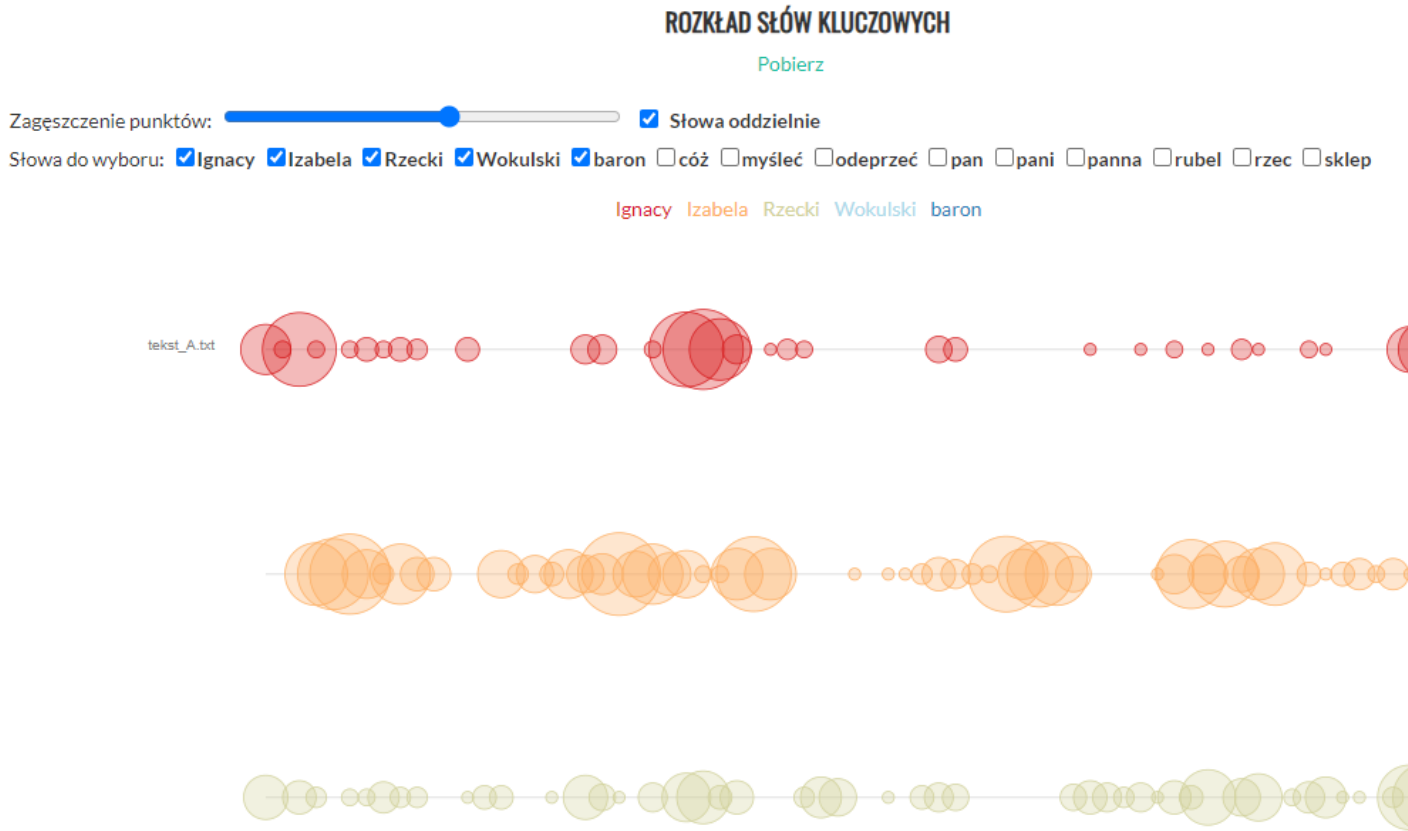
Pobierz

	Forma bazowa	C-value	Liczba wyst
1	tysiąc rubli	185.62	187
2	panna Izabeli	159.57	161
3	pani Stawskiej	90.90	92
4	młody człowiek	64.83	66
5	wielmożny pan	38.67	40
6	pan Ignacy	36.00	37
7	szanowny pan	35.00	37
8	słowo honoru	28.00	28

SŁOWA KLUCZOWE

Pobierz

Ekran statystyk dla korpusu zawierającego tekst “Lalki” Bolesława Prusa.

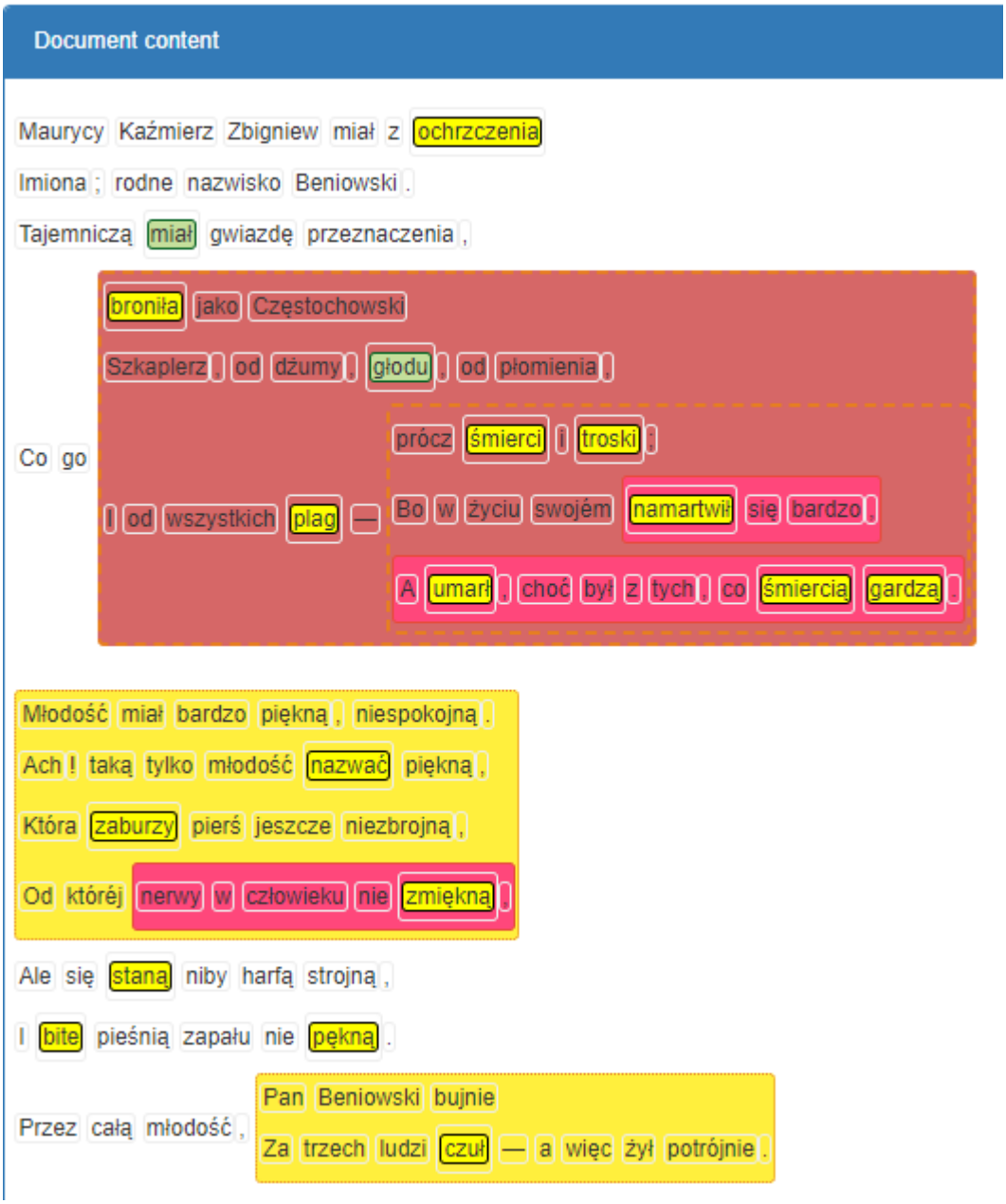


Szczegółowy opis działania serwisu można znaleźć w instrukcji, zaś opis narzędzia można przeczytać w poświęconej mu publikacji.

Inforex – król anotacji

Teksty wgrane do Korpusomatu zostają poddane automatycznej analizie przez narzędzie. Jednak serwis ten nie sprawdzi się, jeśli chcemy nanieść na tekst warstwę własnych anotacji opisujących właściwości językowe, modernizację, cechy gramatyczne bądź inne dowolnie przez nas wybrane informacje dotyczące tekstu i przypisane do poszczególnych jego fragmentów.

Do obsługi anotacji — nanoszonych ręcznie lub automatycznie — przyda się **Inforex**. Jest to narzędzie, które pozwala na stworzenie własnego zestawu anotacji, a następnie nanoszenie ich na tekst. Możliwe jest także zaimportowanie anotacji naniesionych automatycznie np. za pomocą serwisu LEM (o którym więcej za chwilę).



Fragment poematu Juliusza Słowackiego “Beniowski” opisany systemem anotacji

Serwis pozwala też na anotowanie tych samych fragmentów przez więcej niż jednego anotatora, a następnie porównanie i uzgodnienie anotacji. Poszczególne anotacje mogą być grupowane i przeszukiwane, opisywane dodatkowymi notatkami, system umożliwia też wydobycie statystyk dotyczących ich liczebności i dystrybucji. Anotacja pozwala na oznaczenie, a więc później łatwe znalezienie miejsca w tekście, zgrupowanie podobnych fragmentów czy ich zliczenie.

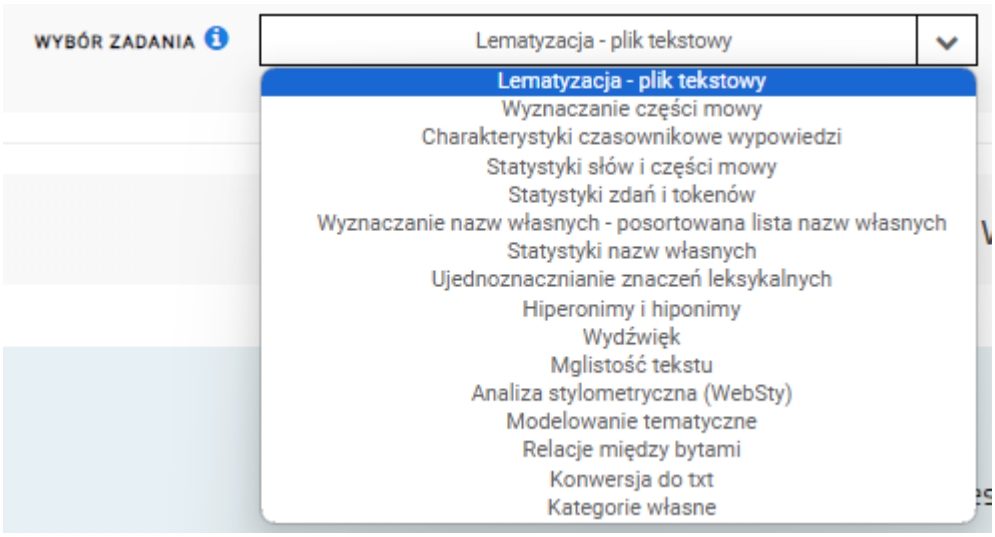
Inforex umożliwia także operacje na wprowadzonym do niego korpusie, takie jak tworzenie list frekwencyjnych, wydobywanie statystyk dotyczących frekwencji i dystrybucji wyrazów czy tagów. Możliwa jest też ręczna korekta tagowania przeprowadzonego automatycznie, co oznacza, że badacz ma znacznie większy

wpływ na jakość uzyskanych wyników i statystyk, niż w przypadku np. Korpusomatu.

Inforex jest narzędziem wszechstronnym, ale w porównaniu z Korpusomatem znacznie mniej przyjaznym w użyciu i intuicyjnym,dlatego warto zapoznać się z instrukcją i materiałami pomocniczymi, aby jak najlepiej wykorzystać jego możliwości.

LEM do eksploracji tekstu

Korpusomat pozwoli nam wydobyć z tekstu podstawowe statystyki i przeszukiwać korpus. Inforex daje nam większe możliwości, ale wciąż są one ograniczone do ściśle określonej listy zadań. Jeśli interesuje nas bardziej zaawansowana, ale też szersza analiza, oraz chcemy móc dostosować ustawienia narzędzia do swoich potrzeb, dobrze sprawdzi się **serwis LEM**, czyli Literacki Eksplorator Maszynowy. Powstał on szczególnie z myślą o literaturoznawcach, ale jego funkcje mogą być wykorzystane do analizy różnych typów tekstów.



Lista zadań wykonywanych przez serwis LEM

LEM integruje wiele pomniejszych usług dostarczanych przez infrastrukturę CLARIN. Część wykonywanych przez niego zadań, takich jak lematyzacja czy wyznaczanie części mowy,może być przydatna przede wszystkimna etapietworzenia materiału do dalszych analiz. Inne mogą same w sobie dostarczyć ciekawych danych, których interpretacja będzie stanowiła zadanie badacza i badaczki. Wśród nich jest charakterystyka czasownikowa wypowiedzi — w efekcie tego działania użytkownik uzyska plik .xlsx, w którym znajdować będą się informacje dotyczące liczby i charakterystyki występowania czasowników (w tym osoba, liczba i rodzaj) oraz bezokoliczników.

Serwis pozwala też na wydobycie statystyk słów i części mowy. Dzięki temu powstaje lista frekwencyjna (czyli uporządkowana lista wszystkich użytych leksemów, ułożona od najliczniej do najmniej licznie występujących), a także lista wszystkich form gramatycznych użytych w całym korpusie. Dzięki odpowiedniej

funkcji możliwe jest też stworzenie listy występujących w korpusie nazw własnych oraz uzyskanie statystyk dotyczących ich występowania.

LEM pozwala także na przeprowadzenie analizy stylometrycznej. Umożliwia ona, w dużym uproszczeniu, znalezienie stylistycznych podobieństw między tekstami. Aby jednak przeprowadzić bardziej zaawansowaną analizę stylometryczną, warto skorzystać z WebSty – narzędzia dającego dostęp do licznych ustawień i pozwalającego dostosować parametry badania ściśle do naszych potrzeb.

Aby szczegółowo zapoznać się z możliwościami, jakie daje LEM, polecam zapoznać się z instrukcją i materiałami pomocniczymi. Szczegóły dotyczące narzędzia opisane są w odpowiedniej publikacji.

Romantycy o miłości

Jak informacje statystyczne mogą zostać wykorzystane w praktyce? Zobaczmy na przykładzie porównania dwóch korpusów: w skład jednego z nich wejdą wszystkie wiersze Juliusza Słowackiego, a w skład drugiego wszystkie wiersze Adama Mickiewicza.

Już na pierwszy rzut oka widać, że korpusy znacząco się różnią:

→**Adam Mickiewicz 160** wierszy, **3952** zdania, **75 036** tokenów,**10 019** leksemów

→**Juliusz Słowacki 220** wierszy, **1588** zdań, **36 345** tokenów,**5485** leksemów

Choć Słowacki napisał więcej pojedynczych wierszy, to łącznie są one o połowę krótsze — zarówno jeśli weźmiemy pod uwagę liczbę zdań, jak i poszczególnych tokenów (czyli najmniejszych jednostek, na które podzielony został tekst) i leksemów. Już ta informacja wskazuje na wyraźne różnice w stylu obu poetów.

Kolejnym punktem analizy może być stworzenie list frekwencyjnych i szukanie ważnych motywów w twórczości obu autorów. Poeci romantyczni w potocznym rozumieniu często kojarzą się z pisaniem o uczuciach, szczególnie o miłości — więc na początek sprawdźmy, jak słownictwo z tego pola semantycznego plasuje się na

liście najczęściej używanych przez dwóch wieszczów słów:

lemat	wiersze JS	wiersze AM	wiersze JS [%]	wiersze AM [%]
miłość	20	28	0,055	0,0373
miłośny	1	4	0,0028	0,0053
miłostka	0	4	0	0,0053
kochać	12	49	0,033	0,0653
kochanka	3	31	0,0083	0,0413
ukochać	3	1	0,0083	0,0013
ukochany	2	2	0,0055	0,0027
kochanek	1	31	0,0028	0,0413
kochanie	0	12	0	0,016
kochany	1	11	0,0028	0,0147

Sama tabela jeszcze nie mówi wiele, dlatego przeanalizujmy całą listę frekwencyjną, aby dowiedzieć się, że:

- u Mickiewicza słowo „miłość” zajmuje 152 miejsce na liście frekwencyjnej. Wyżej znajdują się wyrazy takie jak „człowiek”, „niebo” czy „przyjaciel”. „Miłość” występuje w wierszach Mickiewicza 28 razy, podczas gdy najczęściej używany przez niego rzeczownik — aż 203!
- u Słowackiego „miłość” znajduje się jeszcze dalej na liście frekwencyjnej — zajmuje 352. miejsce, za wyrazami takimi jak np. „listek”, „kościół” czy „świat”. Występuje 20 razy, podczas gdy najczęściej używany rzeczownik ma 171 wystąpień.
-

Skoro nie o miłości, o czym najczęściej pisali romantycy? Choć Mickiewicz pisał, że bardziej przemawiają do niego „czucie i wiara” niż „szkiełko i oko”, to właśnie leksem „oko” jest najczęściej występującym w jego wierszach rzeczownikiem. Na drugim miejscu znajduje się „serce”, na trzecim „świat”, na czwartym „człowiek”, a na piątym „Bóg”. Słowacki najczęściej pisał o „duchu”, a dalej znalazły się słowa „serce”, „człowiek”, „Bóg” i „pan”. Jak widać „serce”, „człowiek” i „Bóg” występują na szczycie listy frekwencyjnej wierszy obu poetów.

Taka — uproszczona z konieczności — analiza pokazuje zarówno istotne podobieństwa, jak i niemożliwe do pominięcia różnice między twórczością Słowackiego i Mickiewicza. A to dopiero wierzchołek góry lodowej! Zachęcam do testowania narzędzi i szukania na własną rękę rozwiązań najlepiej dostosowanych do Waszych potrzeb.

Pierwodruk: Tekst, korpus, wykres. Cyfrowa analiza literatury – przydatne narzędzia, *Otwarta Humanistyka*. (20 czerwca 2024, https://doi.org/10.58079/11ul8)

Linki:

- https://doi.org/10.58079/11ul8

Słowa kluczowe: anotacje, przetwarzanie tekstu, analiza stylometryczna, lematyzacja, korpusomat, Lem, narzędzia cyfrowe

Autor: Anna Mędrzecka-Stefańska