

# Analiza Semantyczna i Ekstrakcja Kolokacji

Dependency Parsing w analizie tekstu literackiego

Przetwarzanie Języka Naturalnego – Zadanie 3

Robert Jacak, Maciej Jamroży

8 lutego 2026

## Streszczenie

Niniejsze sprawozdanie przedstawia analize semantyczna powiesci *Moby Dick* Hermana Melville'a z wykorzystaniem dependency parsing (spaCy). Ekstrakcja kolokacji przymiotnik+rzeczownik i czasownik+rzeczownik pozwoliła na identyfikacje dominujących wzorców językowych tekstu. Wyniki zwizualizowano jako grafy dwudzielne oraz udostepniono przez interaktywne CLI z kolorowa skala częstosci.

## Spis treści

<b>1 Wstęp</b>	<b>3</b>
<b>2 Metodologia</b>	<b>3</b>
2.1 Narzędzia . . . . .	3
2.2 Przetwarzanie tekstu . . . . .	3
2.3 Dependency Parsing . . . . .	3
<b>3 Analiza kolokacji przymiotnikowych</b>	<b>4</b>
3.1 Ranking Top 20 . . . . .	4
3.2 Interpretacja . . . . .	4
<b>4 Analiza kolokacji czasownikowych</b>	<b>5</b>
4.1 Ranking Top 20 . . . . .	5
4.2 Interpretacja . . . . .	5
<b>5 Wizualizacja – grafy dwudzielne</b>	<b>6</b>
5.1 Graf kolokacji przymiotnikowych . . . . .	6
5.2 Graf kolokacji czasownikowych . . . . .	7
<b>6 Interaktywne CLI</b>	<b>7</b>
6.1 Skala kolorów . . . . .	7
6.2 Przykład sesji . . . . .	8
<b>7 Wnioski</b>	<b>8</b>
7.1 Skuteczność metody . . . . .	8
7.2 Ograniczenia . . . . .	9
7.3 Możliwe rozszerzenia . . . . .	9



# 1 Wstep

Celem projektu jest analiza semantyczna tekstu literackiego przy użyciu dependency parsing – metody pozwalającej na identyfikację zależności składniowych między słowami w zdaniu. W odróżnieniu od prostych metod n-gramowych, dependency parsing uwzględnia strukturę gramatyczną, co pozwala na ekstrakcję rzeczywistych kolokacji niezależnie od odległości między słowami.

Jako korpus wybrano powieść *Moby Dick* (1851) Hermana Melville'a, dostępna w domenie publicznej na Project Gutenberg. Tekst liczy 212 796 słów (2802 paragrafów) i reprezentuje bogaty, opisowy styl prozy XIX-wiecznej, obfitującej w rozbudowane konstrukcje przynimotnikowe i złożone relacje czasownik–dopelnienie. Tematyka morska, wielorybowołowa i filozoficzna sprawia, że tekst jest szczególnie interesujący pod kątem analizy kolokacji.

Analiza obejmuje dwa typy relacji:

- **przymiotnik + rzeczownik** (relacja `amod`) – np. *white whale, old man*
- **czasownik + rzeczownik** (relacje `dobj` i `nsubj`) – np. *see whale, man say*

## 2 Metodologia

### 2.1 Narzedzia

- **spaCy** (model `en_core_web_sm`) – biblioteka NLP do tokenizacji, lematyzacji i dependency parsing
- **NetworkX** – biblioteka do tworzenia i analizy grafów
- **Matplotlib** – wizualizacja grafów dwudzielnich
- **colorama** – kolorowe wyjście CLI

### 2.2 Przetwarzanie tekstu

Tekst pobierany jest z Project Gutenberg i zapisywany lokalnie. Następnie usuwane są nagłówki i stopki specyficzne dla Gutenberg (markery **\*\*\* START/END OF THE PROJECT GUTENBERG EBOOK**). Tekst jest normalizowany pod względem białych znaków, z zachowaniem interpunkcji potrzebnej do poprawnego parsowania.

### 2.3 Dependency Parsing

Oczyszczony tekst dzielony jest na paragrafy (separatorem `\n\n`) i przetwarzany wsadowo przez `nlp.pipe()` z rozmiarem batcha 50, co zapewnia wydajną obsługę dużych tekstów.

Dla każdego tokenu analizowana jest relacja syntaktyczna (`dep_`):

- `amod` (adjectival modifier) – token jest przynimotnikiem, jego `head` jest rzeczownikiem. Tworzy parę (przymiotnik, rzeczownik).
- `dobj` (direct object) – token jest rzeczownikiem-dopelnieniem, jego `head` jest czasownikiem. Tworzy parę (czasownik, rzeczownik).

- `nsubj` (nominal subject) – token jest rzeczownikiem-podmiotem (filtrowany dodatkowo przez `pos_ == NOUN`), jego `head` jest czasownikiem. Tworzy parę (czasownik, rzeczownik).

Wszystkie słowa są lematyzowane i konwertowane do małych liter. Filtr `.isalpha()` odrzuca tokeny zawierające znaki niealfabetyczne.

W wyniku analizy wyekstrahowano **12 033** unikalnych par przysłówka+rzeczownik oraz **9 485** unikalnych par czasownik+rzeczownik.

## 3 Analiza kolokacji przysłówkowych

### 3.1 Ranking Top 20

Poniżej przedstawiono 20 najczęściej występujących przysłówków wraz z ich najczęstszymi partnerami-rzeczownikami:

Tabela 1: Top 20 kolokacji przysłówk + rzeczownik

Ranga	Przysłównik	Laczna czest.	Glowni partnerzy
1	old	323	man, time, age
2	other	283	thing, whale, boat
3	great	269	whale, part, leviathan
4	same	173	time, way, instant
5	good	165	time, deal, luck
6	long	164	time, line, row
7	own	163	hand, person, heart
8	such	163	time, thing, man
9	white	145	whale, water, man
10	first	124	time, place, man
11	little	123	man, negro, craft
12	small	117	degree, fish, whale
13	many	115	thing, year, case
14	whole	107	world, life, length
15	last	88	night, year, man
16	large	77	whale, ship, number
17	strange	76	thing, sight, captain
18	wild	73	cry, wind, beast
19	full	72	ship, length, whale
20	poor	70	fellow, devil, lad

### 3.2 Interpretacja

Wyniki doskonale odzwierciedlają charakter tekstu Melville'a:

- **Dominacja *old* (323), *other* (283), *great* (269)** – centralne motywów powieści: starość (Ahab, doświadczeni marynarze – *old man*[84]), wielkość (*great whale*[18]), różnorodność świata morskiego.

- **Kolokacja *white whale*[26]** – najbardziej specyficzna para dla tego korpusu, bezpośrednio identyfikująca tytułowego Moby Dicka. Przymiotnik *white* (145 wystąpień) lączy się też z *water*[8] i *man*[7].
- **Język opisowy** – przymiotniki takie jak *strange* (76), *wild* (73), *poor* (70) budują mroczną, dramatyczną atmosferę powieści (*wild cry*[5], *poor fellow*[11]).
- **Tematyka morska** – rzeczowniki *whale*, *ship*, *fish* pojawiają się jako partnerzy wielu przymiotników.

## 4 Analiza kolokacji czasownikowych

### 4.1 Ranking Top 20

Tabela 2: Top 20 kolokacji czasownik + rzeczownik

Ranga	Czasownik	Laczna czest.	Głowi partnerzy
1	be	1270	whale, thing, man
2	have	489	whale, man, look
3	take	233	turn, place, man
4	make	206	man, passage, whaleman
5	seem	177	man, thing, ship
6	see	174	whale, ye, world
7	give	159	chase, way, order
8	come	139	time, man, darkness
9	go	133	man, whale, ship
10	do	132	bidding, man, ye
11	say	95	ye, word, landlord
12	stand	83	man, harpooneer, tree
13	get	81	chance, peep, start
14	turn	70	head, thunder, face
15	know	69	one, ye, man
16	hold	68	peace, head, mate
17	hear	67	voice, ye, talk
18	keep	64	eye, ship, out
19	find	62	place, doubloon, house
20	show	60	visitor, face, tooth

### 4.2 Interpretacja

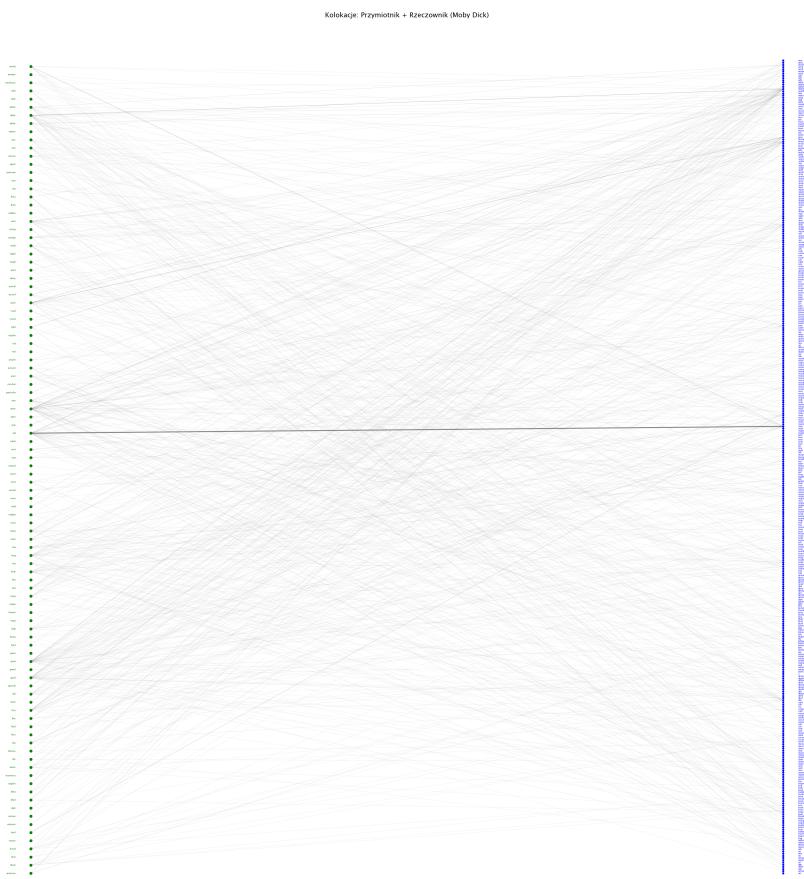
- **be dominuje (1270)** – jako kopula łączy się z podmiotami: *whale*[59], *thing*[31], *man*[26], *ship*[24] – odzwierciedla opisowy, encyklopedyczny styl Melville'a.
- **Czasowniki percepacji** (*see* 174, *seem* 177, *hear* 67) – odzwierciedlają obserwacyjny charakter narracji Ishmaela, który opisuje świat z pozycji świadka.
- **Czasowniki akcji** (*take* 233, *make* 206, *give* 159) – związane z pracą na statku: *give chase*[11], *take turn*[8], *make passage*[5].

- *whale* jako uniwersalny partner – pojawia sie przy czasownikach *be*[59], *have*[13], *see*[12], *go*[6], co potwierdza centralna role wieloryba w fabule.

## 5 Wizualizacja – grafy dwudzielne

Wyniki zostaly zwizualizowane jako grafy dwudzielne, gdzie lewa strona reprezentuje slowa kluczowe (przymiotniki lub czasowniki), a prawa strona – rzeczowniki. Grubosc i przezroczystosc krawedzi sa proporcjonalne do czeostosci kolokacji.

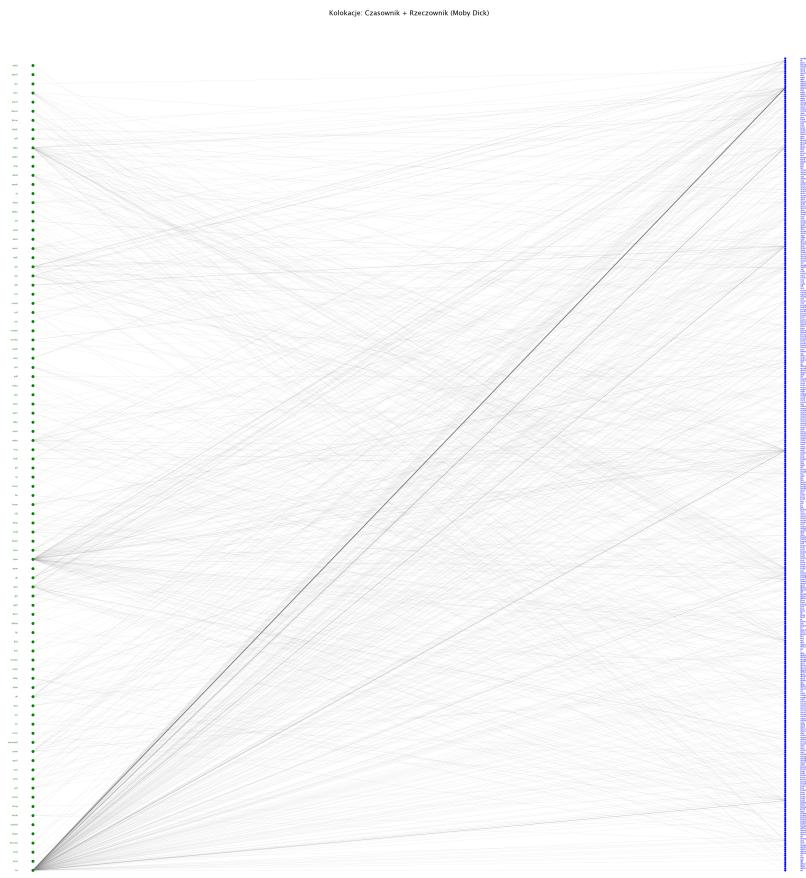
### 5.1 Graf kolokacji przymiotnikowych



Rysunek 1: Graf dwudzielny: przymiotnik + rzeczownik (top 100 przymiotnikow, waga  $\geq 2$ ). Zielone wezly – przymiotniki, niebieskie – rzeczowniki.

Na grafie widoczne sa wyrazne klastry wokol centralnych rzeczownikow: *whale*, *man*, *sea*. Przymiotnik *white* tworzy szczególnie silne polaczenie z *whale*, co odpowiada tytulowemu Moby Dickowi.

## 5.2 Graf kolokacji czasownikowych



Rysunek 2: Graf dwudzielny: czasownik + rzeczownik (top 100 czasownikow, waga  $\geq 2$ ). Zielone wezly – czasowniki, niebieskie – rzeczowniki.

Rzeczowniki *whale* i *man* stanowią centralne wezły grafu, laczace sie z wieloma czasownikami. Wskazuje to na ich dominujaca role w strukturze narracyjnej tekstu.

## 6 Interaktywne CLI

Skrypt udostepnia interaktywne CLI do przeszukiwania wyekstrahowanych kolokacji. Uzytkownik moze:

- Wpisac **2 slowa** – sprawdzenie czestosci konkretnej pary w obu typach relacji
- Wpisac **1 slowo** – wyszukanie wszystkich partnerow danego slowa (jako keyword i jako noun)

### 6.1 Skala kolorow

Wyniki sa kolorowane wedlug czestosci wystapien:

0 wystapien -- CZERWONY (nie wystepuje)  
1 wystapienie -- ZOLTY (rzadkie)

2-10 wystapien -- ZIELONY (umiarkowane)  
>10 wystapien -- NIEBIESKI (czeste)

## 6.2 Przyklad sesji

```
> white whale
Para: 'white' + 'whale'
adj+noun: 26 czeste
verb+noun: 0 nie wystepuje
lacznie: 26 czeste

> whale
Szukam kolokacji dla: 'whale'

Jako rzeczownik (adj+noun) - top 15 przymiotnikow:
white: 26
great: 18
other: 13
large: 7
small: 4

Jako rzeczownik (verb+noun) - top 15 czasownikow:
be: 59
have: 13
see: 12

> old
Szukam kolokacji dla: 'old'

Jako przymiotnik (adj+noun) - top 15:
man: 84
time: 10
age: 8

> quit
Do widzenia!
```

## 7 Wnioski

### 7.1 Skutecznosc metody

Dependency parsing okazal sie znacznie skuteczniejszy niz proste metody n-gramowe w ekstrakcji kolokacji:

- **Niezaleznosc od odleglosci** – relacja amod laczy przymiotnik z rzeczownikiem nawet gdy miedzy nimi znajdują się inne słowa (np. “*the great white sperm whale*”).
- **Rozroznianie rol syntaktycznych** – oddzielenie podmiotu (**nsubj**) od dopełnienia (**dobj**) daje bogatszy obraz relacji.

- **Lematyzacja** – sprowadzenie slow do form bazowych (np. “*whales*” → “*whale*”) konsoliduje statystyki.

## 7.2 Ograniczenia

- Model `en_core_web_sm` to maly model – wieksze modele (np. `en_core_web_lg`) moga dawac dokladniejsze parsowanie.
- Analiza nie uwzglednia negacji ani modalnosci – “*no white whale*” jest liczone tak samo jak “*white whale*”.
- Archaiczny jezyk XIX-wieczny moze powodowac bledy w parsowaniu wspolczesnym modelem NLP.
- Tekst w jezyku angielskim – analiza polskojezycznego tekstu wymagalaby innego modelu.

## 7.3 Mozliwe rozszerzenia

- Analiza kolokacji z uwzglednieniem rozdzialow – sledzenie zmian kolokacji w toku narracji.
- Porownanie z innymi powiesciami tego samego okresu.
- Wykorzystanie wiekszych modeli spaCy lub transformerow (np. spaCy z backendem transformer).
- Analiza sentymetu kolokacji – czy dane pary pojawiaja sie w kontekstach pozytywnych czy negatywnych.

## 8 Bibliografia

1. spaCy – Industrial-Strength Natural Language Processing. <https://spacy.io/>
2. Project Gutenberg – *Moby Dick* by Herman Melville. <https://www.gutenberg.org/ebooks/2701>
3. NetworkX – Network Analysis in Python. <https://networkx.org/>
4. Matplotlib – Visualization with Python. <https://matplotlib.org/>
5. Melville, H. (1851). *Moby-Dick; or, The Whale*. Harper & Brothers.