2025/2026

Analiza Statystyczna Korpusu Języka Fińskiego

Przetwarzanie języka naturalnego w systemach sztucznej inteligencji

Robert Jacak, Maciej Jamroży

Spis treści

1. Wstęp

1.1 Cel projektu

Celem niniejszej pracy jest zbadanie struktury leksykalnej języka fińskiego przy użyciu metod lingwistyki komputerowej. Projekt koncentruje się na weryfikacji prawa Zipfa, określeniu minimalnego słownictwa niezbędnego do rozumienia tekstu oraz identyfikacji "rdzenia języka" poprzez analizę grafową, a także określenie 50 najczęstszych rzeczowników.

1.2 Materiał badawczy i narzędzia

Analizie poddano zrzut bazy danych fińskiej Wikipedii (fiwiki-latest-pages-articles.xml.bz2).

Proces przetwarzania obejmował:

- **Ekstrakcję treści z XML:** Gensim WikiCorpus.
- **Lematyzację:** Biblioteka Simplemma.
- **Analizę morfosyntaktyczną:** Model neuronowy Stanza dla języka fińskiego.
- **Tłumaczenie na język polski:** Deep-translator

1.3 Statystyki korpusu

- **Całkowita liczba tokenów:** 135 064 600
- **Liczba unikalnych wyrazów (lematów):** 2 238 241

2. Analiza Rozkładu Częstotliwości (Prawo Zipfa)

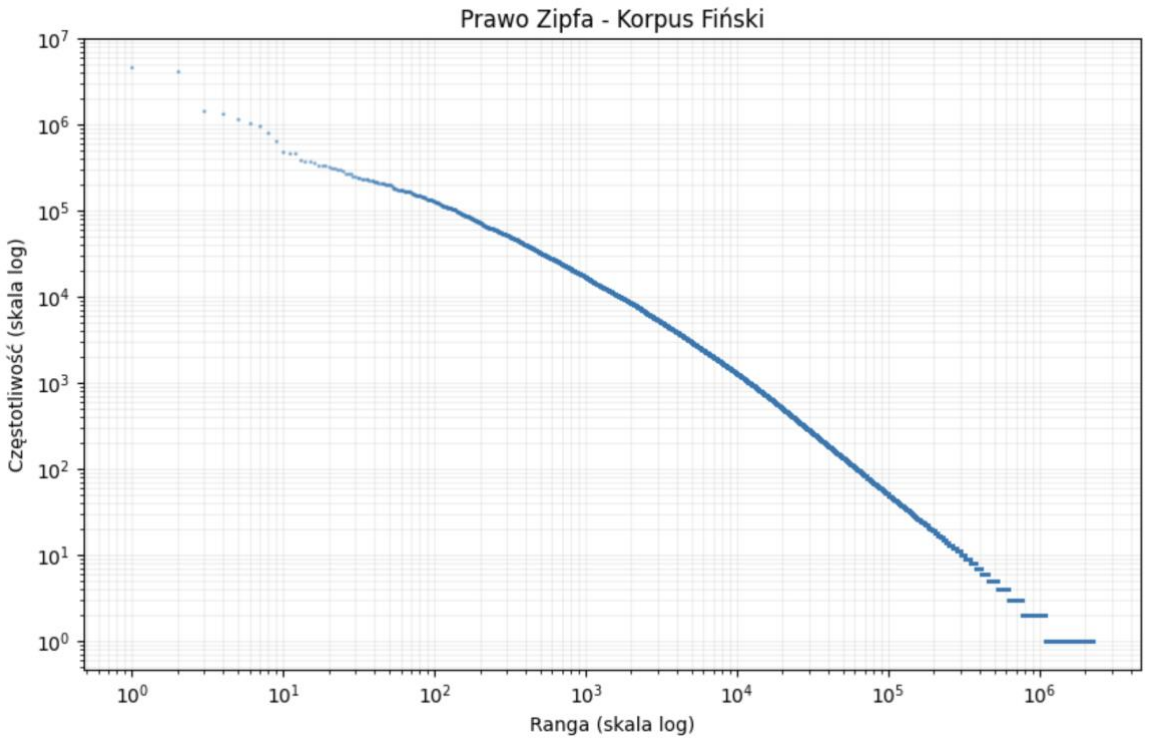
2.1 Prawo Zipfa

Prawo Zipfa to empiryczne prawo statystyczne opisujące rozkład częstości słów w języku naturalnym.

Częstotliwość występowania słowa jest odwrotnie proporcjonalna do jego rangi w rankingu częstości.

$$f(r) \propto \frac{1}{r}$$

Wygenerowany wykres w skali logarytmicznej przedstawia zależność między rangą słowa a jego częstotliwością w badanym korpusie języka fińskiego.



Wnioski:

Krzywa wykazuje charakterystyczny, liniowy trend zgodny ze wzorem $f(r) \propto \frac{1}{r}$.

Potwierdza to, że język fiński, mimo swojej złożonej morfologii, poddaje się uniwersalnym prawom statystycznym języka.

Intuicyjnie: słowo o randze 1 występuje dwa razy częściej niż słowo o randze 2, trzy razy częściej niż słowo o randze 3 itd.

2.1 Najczęstsze słowa

Ranga	Słowo	Liczba wystąpień	Tłumaczenie
1	ja	4678425	I
2	olla	4269307	Być
3	luokka	1474868	klasa
4	hän	1366437	On
5	vuonna	1155135	w roku
6	se	1049938	To
7	joka	972321	Który
8	vuosi	807959	rok
9	myös	647121	Również
10	lähde	486450	źródło
11	ei	471196	NIE
12	suomi	469295	fiński

13	saada	395468	Dostawać
14	sekä	374602	mieszany
15	muu	372869	Inny
16	the	357427	the
17	mutta	341328	Ale
18	ensimmäinen	332906	Pierwszy
19	luku	332463	numer
20	alue	319424	obszar

Najczęstsze słowa to głównie słowa funkcyjne (spójniki, przyimki, zaimki) oraz czasowniki posiłkowe, co jest typowe dla korpusów języków naturalnych. Słowa treściowe (rzeczowniki) pojawiają się dopiero na dalszych pozycjach.

3. Pokrycie Tekstowe (Zrozumienie Języka)

Jaki procent słownictwa trzeba znać, aby zrozumieć dany procent tekstu?

Pokrycie tekstu	Wymagana liczba słów	Procent słownictwa
50%	814	0.0364
80%	10937	0.4886
90%	43680	1.9515
95%	146706	6.5545
99%	982731	43.9064

Wnioski: Wykładniczy wzrost liczby wymaganych słów dla wyższych progów pokrycia wskazuje na tzw. "długi ogon" (Long Tail). Aby przejść z poziomu rozumienia 80% do 90%, należy niemal czterokrotnie zwiększyć zasób słownictwa

4. Rdzeń Języka – Analiza Współwystępowania

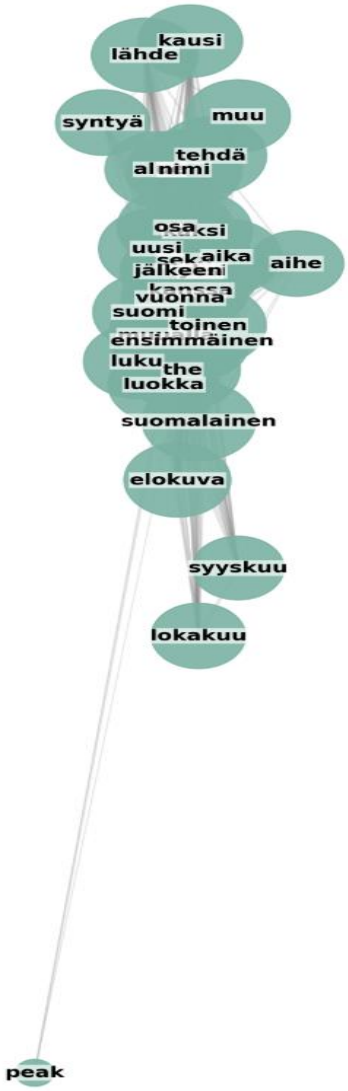
Rdzeń języka zidentyfikowano poprzez analizę **współwystępowania słów** (co-occurrence), modelując tekst jako graf nieskierowany, w którym węzłami są najczęstsze słowa znaczące, a krawędziami – ich bezpośrednie sąsiedztwo w zdaniach. Jako miarę ważności przyjęto **centralność stopnia** (Degree Centrality), czyli liczbę unikalnych słów, z którymi dany wyraz tworzy relacje. Słowa o najwyższej centralności, określane jako **Hub Words**, stanowią "szkielet" semantyczny języka, łącząc różnorodne konteksty tematyczne i zapewniając spójność tekstu.

4.1 Słowa Węzłowe

Poniższa tabela przedstawia słowa, które najczęściej łączą się z innymi w analizowanych zdaniach, czyli mają największa centralność.

Ranga	Słowo	Liczba sąsiadów	Centralność znorm.	Tłumaczenie
1	sekä	28	0.9655	mieszany
2	vuonna	28	0.9655	w roku
3	yksi	28	0.9655	jeden
4	osa	27	0.931	część
5	nimi	27	0.931	nazwa
6	toinen	27	0.931	drugi
7	ensimmäinen	27	0.931	Pierwszy
8	uusi	27	0.931	nowy
9	the	27	0.931	the
10	kanssa	27	0.931	z
11	kaksi	27	0.931	dwa
12	muualla	27	0.931	gdzie indziej
13	jälkeen	27	0.931	Po
14	suomalainen	26	0.8966	fiński
15	luokka	25	0.8621	klasa
16	suomi	25	0.8621	fiński
17	tehdä	25	0.8621	robić
18	alue	24	0.8276	obszar
19	aika	24	0.8276	czas
20	elokuva	23	0.7931	film
21	luku	23	0.7931	numer
22	lähde	23	0.7931	źródło
23	kausi	22	0.7586	sezon
24	muu	22	0.7586	Inny
25	lokakuu	18	0.6207	Październik
26	syntyä	18	0.6207	urodzić się
27	aihe	18	0.6207	temat
28	syyskuu	17	0.5862	Wrzesień
29	peak	3	0.1034	Szczyt
30	lempo	0	0.0	Ładny

4.2 Wizualizacja



Graf współwystępowania 30 najważniejszych słów fińskich. Rozmiar węzła odpowiada liczbie sąsiadów.

5. Analiza Semantyczna – Najczęstsze Rzeczowniki

Przy użyciu modelu neuronowego Stanza wyekstrahowano oraz przetłumaczono 50 najczęstszych rzeczowników. Stanowią one tematyczny szkielet korpusu.

lemma	liczba wystąpień	tłumaczenie
vuosi	10638	rok
suomi	3481	fiński
luku	3287	numer
tammikuu	3151	Styczeń
alue	2990	obszar

nimi	2509	nazwa
kunta	2419	miasto
maa	2408	kraj
kalenteri	2366	kalendarz
Suomi	2361	fiński
osa	2342	część
aika	2278	czas
kaupunki	2125	miasto
päivä	1845	dzień
yhdysvallat	1809	Stany Zjednoczone
kieli	1743	język
kuningas	1669	król
eurooppa	1604	Europa
helmikuu	1568	Luty
luokka	1543	klasa
ihminen	1543	Człowiek
maaliskuu	1411	marzec
sota	1358	wojna
ruotsi	1338	szwedzki
historia	1296	historia
kiina	1285	Chiny
käyttö	1205	używać
etelä	1199	południe
ranska	1190	Francja
tapahtuma	1186	wydarzenie
joukko	1087	ustawić
venäjä	1067	Rosja
huhtikuu	1039	Kwiecień
sana	1030	słowo
maailma	1029	świat
helsinki	1028	Helsinki
pinta	1027	powierzchnia
saksa	1006	Niemcy
tiedosto	982	plik
toukokuu	979	Móc
keisari	972	cesarz
valtio	966	państwo
pohjois	951	północ
saari	935	wyspa
asukas	923	rezydent
maa#kunta	923	województwo
presidentti	895	prezydent

alku	858	wstępny
itä	833	wschód
de	833	de

Wśród 50 najczęstszych rzeczowników dominują kategorie takie jak:

1. **Czas:** *vuosi – rok, päivä – dzień,* tammikuu – styczeń.
2. **Geografia:** *alue – obszar, kunta – miasto,* yhdysvallat – Stany Zjednoczone.
3. **Społeczeństwo/Ludzie:** *ihminen - człowiek, keisari - cesarz,* presidentti – prezydent.

6. Podsumowanie

6.1 Wnioski

Analiza korpusu języka fińskiego pozwoliła na:

1. Potwierdzenie uniwersalności prawa Zipfa także dla języka fińskiego.
2. Wyznaczenie progu leksykalnego: znajomość **814** słów pozwala na zrozumienie **50%** tekstu, co stanowi "statystyczny rdzeń języka", żeby zrozumieć **90%** tekstu należy znać znacznie więcej słów – aż **43680**.
3. Zidentyfikowanie "rdzenia strukturalnego" metodą grafową.

6.2 Ograniczenia badania

Największym ograniczeniem tego projektu jest to, że opiera się on wyłącznie na Wikipedii. Jest to encyklopedia, a nie zapis naturalnej rozmowy, dlatego słowa techniczne (takie jak „kategoria” czy „źródło”) pojawiają się na szczycie list o wiele częściej, niż w normalnym życiu. Dodatkowym wyzwaniem jest trudna budowa języka fińskiego, który „skleja” wiele słów w jedno. Programy komputerowe mogą mieć problem z bezbłędnym sprowadzeniem ich do formy podstawowej, co mogło wpłynąć na dokładność niektórych wyliczeń.

7. Struktura projektu

- **`1_process_corpus.py`** – Przetwarza surowy zrzut XML Wikipedii i wykonuje lematyzację całego korpusu do formatu CSV.
- **`2_analyze_zipf.py`** – Generuje wykres logarytmiczny weryfikujący prawo Zipfa oraz oblicza progi pokrycia tekstu (zrozumienie 50-99%).
- **`3_extract_nouns.py`** – Wykorzystuje model neuronowy Stanza do ekstrakcji i zliczenia najczęstszych rzeczowników oraz nazw własnych.
- **`4_analyze_core.py`** – Identyfikuje "rdzeń języka" poprzez budowę grafu współwystępowania słów i obliczenie ich centralności (Degree Centrality).
- **`5_translate_nouns.py`** – Automatycznie tłumaczy wyekstrahowane rzeczowniki na język polski przy użyciu API Google Translate.
- **`6_translate_core.py`** – Tłumaczy zidentyfikowane słowa węzłowe (Hub Words) w celu ich interpretacji semantycznej w raporcie.