Koherencja i prawo Zipfa

Rafał Gęgotek
Kamil Słowiński
Krzysztof Michalski
Dariósz Nostkiewicz

Co to jest Koherencja w lingwistyce

Koherencja to tzw. spójność globalna, jest efektem połączenia wszystkich znaczeń w komunikacie z wiedzą o świecie odbiorców (aluzja, dowcip, intertekstualność).

Gdy zasób wiedzy nadawcy i odbiorcy pokrywa się dopuszczalna jest wówczas w komunikacji spójność globalna (tematyczna).

Jest charakterystyczna dla monologów, tekstów artystycznych.

Pojęciem tym posługują się językoznawcy zajmujący się stylistyką lingwistyczną.

Koherencja przykład

Zrozumienie spójnego dyskursu obejmuje nie tylko indywidualną interpretację każdej wypowiedzi, ale także sposób, w jaki te wypowiedzi odnoszą się do siebie.

Szereg spójnych stwierdzeń można sformułować tak, jak pokazano na kolejnych slajdach:

Koherencja przykład

- **Paul je jabłko. Bardzo lubi owoce.**

W powyższym przykładzie możemy zrozumieć, że badany Paul je **jabłko**, ponieważ lubi **owoce**. Stwierdzenia te straciłyby swoją relację spójności, gdyby obiekt ***jabłko*** został zastąpiony czymś innym, co nie jest owocem, jak na następnym slajdzie.

Koherencja przykład

- **Paul je brokuły. Bardzo lubi owoce.**

Podczas gdy pojęcie spójności ma pożytek z utrzymywania wiarygodnego następstwa zdań w przemówieniu, pojęcie spójności pozwala na zestawienie tych stwierdzeń w jasny i logiczny sposób. Aby to zrobić, musi wziąć pod uwagę różne kontekstualne aspekty omawianego dyskursu. Znajdziemy między innymi wiedzę o świecie, jaką posiadają osoby zaangażowane w przemówienie (mówca (mówcy) i rozmówca (rozmówcy)). W tym sensie można przypuszczać istnienie dużego zbioru wiedzy relacyjnej, pozwalającej wyjaśnić związek między dwoma proponowanymi stwierdzeniami.

Koherencja przykład

- **Paul tańczy. On jest szczęśliwy.**

W powyższym przykładzie czytelnik raczej zakwestionuje spójność relacji między faktem, że Paweł tańczy, a faktem, że jest szczęśliwy. Jednak stwierdzenia mogą być spójne, jeśli przyjmie się, że istnieje logiczny związek w sytuacji kontekstualnej między radością Pawła a tym, dlaczego tańczy. Wydaje się, że zjawisko to pokazuje, że potrzeba uwiarygodnienia spójności między dwoma lub więcej stwierdzeniami jest naturalną częścią naszej zdolności rozumienia języka.

Formy relacji koherencyjnych

Różne relacje reprezentowane przez pojęcie **spójności** mogą objawiać się w kilku formach. Jednak dokładny charakter większości tych relacji nigdy nie został jasno określony, ponieważ wielu autorów nadal czyni je przedmiotem debaty. Najczęściej obserwowanymi formami **relacji koherencyjnych** są: **narracja** (lub czasowość), **przyczynowość** (związki przyczynowo-skutkowe), **podobieństwo** i **dopracowanie**.

Formy relacji koherencyjnych

Narracja (sekwencja czasowa): pierwsze zdanie zdania stanowi zdarzenie poprzedzające drugie zdanie.

- Paul się napił. Zjadł jabłko.

Przyczynowość (skutek, stosunek przyczyny i skutku): drugie zdanie stanowi zdarzenie, które logicznie następuje po pierwszym zdaniu. Dlatego drugie zdanie musi mieć miejsce, jeśli występuje pierwsze.

- Paul uderzył Charlesa. On cierpiał.

Formy relacji koherencyjnych

Podobieństwo (relacja równoległa): Zdania składające się na zdanie są podobne.

- Paul walczył z Charlesem. Pierre walczył z Jean.

Opracowanie: Drugie zdanie stanowi podzbiór pierwszego zdania. W tym sensie drugie zdanie jest bardziej szczegółowe, bardziej szczegółowe niż pierwsze, tworząc w nim pewną ciągłość.

- Paul poszedł do restauracji. Zjadł stek.

Co to jest prawo Zipfa

Prawo Zipfa – prawo empiryczne głoszące, że wiele rodzajów danych tworzonych przez ludzi lub odnoszących się do ich zachowań cechuje charakterystyczny rozkład wartości, w którym dystrybucja częstotliwości występowania poszczególnych wartości jest odwrotnie proporcjonalna do ich rangi statystycznej.

Pod koniec **XIX wieku** francuski stenograf i leksykograf **Jean-Baptiste Estoup**, badając zasady stenografii, ustalił podstawowe zasady statystyczne dotyczące tekstu. Twierdzenia francuskiego badacza zweryfikował i uściślił amerykański lingwista **George Kingsley Zipf**.

Prawo Zipfa dla języków naturalnych

Pierwotnie prawo to zostało sformułowane dla języków naturalnych, w których zaobserwowano, że gdy na podstawie ich korpusów językowych ustali się wykaz wyrazów ułożonych w malejącym porządku częstotliwości ich występowania, to ranga (numer porządkowy) wyrazu jest odwrotnie proporcjonalna do częstotliwości, zatem iloczyn częstotliwości i rangi powinien być wielkością stałą.

Przykładowo: w korpusie Browna dla języka angielskiego w wersji amerykańskiej, najczęściej występujące słowo **"the"** stanowi aż **7%** wszystkich słów, drugie w kolejności **"of"** stanowi **3,5%**, trzecie **"a" 1,75%**, zaś pierwsze **135 słów** składa się na **50%** objętości całego korpusu.

Prawo Zipfa dla języków naturalnych

Matematycznie można to wyrazić w formie równania:

r * f = constans,

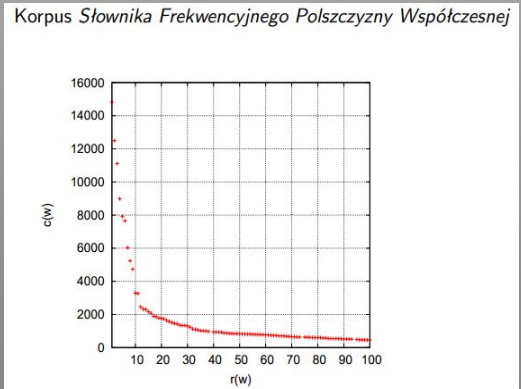
gdzie **r** jest to ranga wyrazu w tekście lub grupie tekstów, a **f** częstotliwość jego występowania.

W odpowiednio obszernych korpusach językowych wartość stałej jest charakterystyczna dla danego języka, a prawo jest spełnione niemal doskonale dla pierwszych, najczęściej występujących 200-300 słów. W poszczególnych tekstach zależy ona natomiast od stylu i tematyki. Porównanie rozkładu Zipfa obliczonego dla korpusu języka z rozkładem dla danego tekstu pozwala na ocenę stylu autora i jego zrozumiałość przez przeciętnego czytelnika. Czym bardziej rozkład dla analizowanego tekstu jest zgodny z rozkładem ogólnym dla języka, w którym go napisano, tym jest on bardziej zrozumiały dla większości osób posługujących się na co dzień tym językiem.

Prawo Zipfa objaśnienia

Korpus *Słownika Frekwencyjnego Polszczyzny Współczesnej*

ranga **r(w)**	częstość **c(w)**	słowo **w**
1	14767	w
2	12473	i
3	11093	się
...	...	...



- Najczęstsze słowo ma rangę równą **1**, drugie co do częstości ma rangę równą **2**, trzecie co do częstości ma rangę równą **3**, ... itd,
- Im większa ranga, tym mniejsza częstość,
- Definicja nie mówi, jak szybko ranga maleje z częstością,
- Jeżeli słowo **x** ma rangę **10** razy większą niż słowo **y**, to słowo **x** ma częstość **10** razy mniejszą niż słowo **y**,
- Częstotliwość słowa jest odwrotnie proporcjonalna do jego rangi:
  - **f = constans / r**
- Prawo Zipfa jest najsłynniejszym ilościowym prawem językowym,
- Badaniem ilościowych praw językowych zajmuje się lingwistyka kwantytatywna,
- Rozkłady ranga-częstość podobne do prawa Zipfa pojawiają się także poza lingwistyką:
  - rozkład cytowań artykułów naukowych (prawo Lotki),
  - rozkład dochodów ludności (prawo Pareto, zasada 80/20),
  - rozkład wielkości miast (prawo Gibrata).

Koherencja i prawo Zipfa

Prawo **Zipfa** nie obowiązuje dla podzbiorów lub sumy zbiorów **Zipfa**. W przypadku **podzbiorów**, niektóre **brakujące elementy** nieuchronnie powodują odchylenia od czystego prawa **Zipfa** w podzbiorze, zwłaszcza gdy te **„dziury**” występują dla największych elementów oryginalnego zestawu, przy czym ten problem ma kluczowe znaczenie dla wiodących elementów zestawu. Podobnie połączenie lub agregacja zestawów **Zipfa** nie dziedziczy właściwości koherencji zestawów oryginalnych, ponieważ **repliki** lub **elementy o bardzo podobnej wielkości niszczą integrację w zestawach zagregowanych.**

Koherencja i prawo Zipfa

Jeśli weźmiemy pod uwagę dwie repliki tego samego zestawu, to połączenie dwóch replik nie może być opisane przez to samo prawo.

Takie elementarne przykłady pokazują kluczową rolę, jaką odgrywa właściwość wewnętrznej spójności lub kompletności całego badanego zbioru, którą nazywamy „**koherencją**".