

Nad wyraz ważne przecinki i kropki

13 grudnia 2016

Nasze języki są strukturami bardziej kompletnymi, niż się nam dotąd wydawało? – Tak! Wynika to z analiz częstotliwości występowania wyrazów w najśłynniejszych dziełach literatury. Naukowcy z Krakowa pokazali, że w takich analizach nad wyraz ważne są... znaki interpunkcyjne.

Z dobrym przybliżeniem można powiedzieć, że za 80 proc. naszych sukcesów – odpowiada 20 proc. naszych starań. Ta słynna zależność sprawdza się w zaskakująco wielu dziedzinach. Widać ją na przykład we wszystkich językach, zarówno w mowie, jak i w piśmie: za 80 proc. długości wypowiedzi odpowiada zaledwie 20 proc. najpopularniejszych słów. Opisu ją to zjawisko zależność była jednym z pierwszych poznanych praw potęgowych i dziś nosi nazwę prawa Zipfa.

Szybko się okazało, że w rzeczywistości nie jest ona tak banalna, jakby mogłoby się na pierwszy rzut oka wydawać. Naukowcy z Instytutu Fizyki Jądrowej Polskiej Akademii Nauk (IFJ PAN) w Krakowie właśnie wykazali, że pewne zagadkowe cechy prawa Zipfa, od dziesięcioleci intrygujące zwłaszcza osoby zajmujące się statystyczną analizą tekstów literackich, są konsekwencją niedoceniania roli jednego ze składników języka – interpunkcji. O badaniach poinformowali przedstawiciele IFJ PAN w przesłanym PAP komunikacie.

KSIAŻKOWE FRAKTALE I SŁÓW GIĘCIE

„Mniej więcej rok temu za pomocą szczegółowych analiz statystycznych wykazaliśmy, że długości zdań w literaturze, czyli odległości między kropkami, wykazują bardzo złożone zależności o charakterze multifraktałnym, szczególnie wyraźnie widoczne w dziełach należących do gatunku określanego jako nurt

świadomości

(<http://naukawpolsce.pap.pl/aktualnosci/news,408115,fizycy-odkryli-multifraktale-min-w-dzielach-jamesa-joycea.html>). Był to intrygujący rezultat, który skłonił nas do przyjrzenia się z większą uwagą roli innych znaków interpunkcyjnych, zwłaszcza w kontekście prawa Zipfa. Wyniki, które teraz otrzymaliśmy, każą nam w nowy sposób spojrzeć nie tylko na rolę interpunkcji w języku, ale nawet na sam język” – mówi prof. Stanisław Drożdż z IFJ PAN i Politechniki Krakowskiej.

Wykresy obrazujące prawo Zipfa w przypadku tekstów wykonuje się według mało skomplikowanej procedury. Dla każdego wyrazu zlicza się, jak często występuje on w tekście. Temu wyrazowi, który występuje najczęściej, przyporządkowuje się rząd 1, kolejnemu – rząd 2 itd. Prawo Zipfa mówi, że prawdopodobieństwo wystąpienia wyrazu jest odwrotnie proporcjonalne do jego rzędu: im większy rząd, tym prawdopodobieństwo to jest mniejsze. Wykresy przedstawiające tę zależność mają (na skali logarytmicznej) postać prostej.

Od czasu spopularyzowania przez amerykańskiego lingwistę George’a Zipfa, jego prawo nie przestaje zadziwiać. Jak coś tak skomplikowanego, jak struktury tworzone za pomocą języka, można opisać tak prostym prawem? Zagadek było więcej. Dość wcześnie zauważano, że wykresy dotyczące częstotliwości występowania wyrazów dla rzędów bliskich jedności odginają się nieco w dół od prostej.

Owo odgięcie szczególnie intrygowało Benoit Mandelbrota, wielkiego francuskiego matematyka, który zajmował się tym zagadnieniem przez wiele lat. Zaproponował on nawet własną poprawkę do oryginalnego prawa Zipfa, lepiej odwzorowującą odgięcie (warto w tym miejscu wspomnieć, że to m.in. prace nad prawem Zipfa pchnęły Mandelbrota ku koncepcji fraktali).

NIE SZLI NA SKRÓTY

W swoich najnowszych badaniach fizycy z IFJ PAN poddali analizie teksty napisane w sześciu językach indoeuropejskich,

należących do grup germańskiej (angielski i niemiecki), romańskiej (francuski i włoski) i słowiańskiej (polski i rosyjski). Użyto dzieł literackich z archiwów Projektu Gutenberg (www.gutenberg.org). Wszędzie usunięto kropki po skróconych wyrazach (np. w skrócie „np.”), zlikwidowano przypisy, numery stron i inne znaki o charakterze bardziej typograficznym.

„Ostatecznie znakami interpunkcyjnymi uwzględnianymi w naszych analizach były przecinki, dwukropki i średniki, a za znaki kończące zdania uznaliśmy kropki, wykrzykniki, pytajniki i wielokropki” – precyzuje dr hab. Jarosław Kwapien (IFJ PAN), jeden ze współautorów publikacji w „Information Sciences”.

Wśród badanych dzieł znalazły się m.in.: „1984” George’a Orwella, „Moby Dick” Hermana Melville’a, „Ulisses” Jamesa Joyce’a, „Podróż Guliwera” Jonathana Swifta, „Przeminęło z wiatrem” Margaret Mitchell, „Tako rzecze Zaratustra” Friedricha Nietzschego, „Proces” Franza Kafki, „Czarodziejska góra” Thomasa Manna, „Madame Bovary” Gustave’a Flauberta, „Upiór w operze” Gastona Leroux, „Wahadło Foucault” Umberto Eco, „Przedwiośnie” Stefana Żeromskiego, „Ziemia obiecana” Władysława Reymonta, „Lalka” Bolesława Prusa, „Anna Karenina” i „Wojna i pokój” Lwa Tołstoja, a także „Bracia Karamazow” Fiodora Dostojewskiego.

INTERPUNKCJA SIĘ LICZY!

Uwzględnienie znaków interpunkcyjnych doprowadziło do interesującego rezultatu: zakrzywienie wykresu Zipfa dla wyrazów o rzędach bliskich jedności praktycznie zniknęło. Nowe wykresy układały się niemal dokładnie wzdłuż prostej, a więc zgodnie z pierwotną postacią prawa Zipfa. Poprawka do tego prawa, wprowadzona przez Mandelbrota, okazała się w ogóle zbędna.

„Gdy znaki interpunkcyjne zaczynamy traktować jak wyrazy, to one zaczynają zajmować rzędy bliższe jedności i zakrzywienie

wykresów Zipfa w zasadzie znika. Zatem po uwzględnieniu interpunkcji każdy badany przez nas język zaczyna wyglądać jak twór bardziej spójny! Dlatego uzasadnione wydaje się stwierdzenie, że interpunkcja jest dla języka równie ważna jak słowa, a język bez niej staje się po prostu niekompletny” – mówi prof. Drożdż.

TO CIĄGŁE NIE JEST TAKIE PROSTE

Nowe wykresy nadal wykazują kilka interesujących cech. Na przykład w językach słowiańskich po uwzględnieniu znaków interpunkcyjnych wykres Zipfa układa się praktycznie idealnie wzdłuż prostej, podczas gdy w czterech pozostałych co prawda wyraźnie się prostuje, lecz w nieco mniejszym stopniu i pewne odchylenie wciąż jest widoczne, zwłaszcza w języku angielskim.

„Może w trakcie analiz tekstów w językach niesłowiańskich nie uwzględniliśmy ich jakichś dodatkowych, specyficznych cech?” – zastanawia się prof. Drożdż. Rozważa też inną, ciekawszą interpretację: „Może być też tak, że przyczyna niepełnej redukcji zakrzywienia tkwi w samym języku. Na przykład w angielskim może mieć źródło w łatwo dostrzegalnej tendencji autorów do ograniczania liczby znaków interpunkcyjnych. Jeśli ten ostatni powód jest prawdziwy, może warto się zastanowić, czy nadmierne redukovanie interpunkcji na pewno jest działaniem korzystnym i czy czasami nie niszczy wewnętrznej integralności języka?” – pyta badacz.

Odkrycie fizyków z IFJ PAN potencjalnie może mieć znaczenie wykraczające poza badania lingwistyczne. Odchylenie części wykresu Zipfa od prostej dla rzędów bliskich jedności jest obserwowane w wielu dziedzinach i ma różnorodną, nie zawsze do końca dobrze zrozumianą naturę. Skoro w wykresach przygotowanych na podstawie dzieł literackich zniknęło ono po uwzględnieniu czynnika powszechnego, lecz dotychczas uważanego za nieistotny, być może także w niektórych innych przypadkach można byłoby je usunąć poprzez uwzględnienie w analizach elementu do tej pory uznawanego za pozbawiony większej roli.

Źródło: NaukawPolsce.PAP.pl