



Kraków, 30 listopada 2016

Kluczowa rola pomijanego składnika języka i literatury

Wszystko wskazuje na to, że udało się rozwiązać zagadkę występującą we wszystkich ludzkich językach, od niemal stu lat intrygującą zwłaszcza matematyków i lingwistów zajmujących się badaniami statystycznymi literatury. Nowe analizy częstotliwości występowania wyrazów w najsłynniejszych książkach, przeprowadzone przez naukowców z Instytutu Fizyki Jądrowej PAN w Krakowie, pokazują, że nasze języki są strukturami bardziej złożonymi i bardziej kompletnymi niż się nam dotychczas wydawało.

Z dobrym przybliżeniem, za 80% twoich sukcesów odpowiada 20% spośród twoich 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% długości wypowiedzi odpowiada zaledwie 20% najpopularniejszych słów, co prawdopodobnie wynika m.in. z faktu, że gdy rozmawiamy ze sobą, chcemy przekazać jak najwięcej treści, ale jak najmniejszym wysiłkiem. Opisująca 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 niedoceniań roli jednego ze składników języka.

„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 multifrakalnym, szczególnie wyraźnie widoczne w dziełach należących do gatunku określanego jako nurt świadomości. 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. dr hab. Stanisław Drożdż (IFJ PAN, Politechnika Krakowska).

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. (w bogatych tekstach rzędy mogą przekraczać 10 tys.; wyrazy egzotyczne zwykle pojawiają się powyżej rzędu 1000). 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 polskiego pochodzenia, 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).

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), o długości co najmniej pięciu tysięcy zdań. W każdym z języków z co najmniej pięciu różnych tekstów utworzono także jeden dodatkowy, liczący około miliona wyrazów. Wszędzie usunięto wyrazy niezwiązane z przenoszoną treścią, takie jak „rozdział”, „część” czy „epilog”, a także charakterystyczne dla danego języka skróty, np. „Mr.” czy „Dr.”, które po wykasowaniu kropek traktowano jako odrębne wyrazy. Zlikwidowano także przypisy, numery stron i inne znaki o charakterze bardziej typograficznym: cudzysłowy i nawiasy.

„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 znanym czasopiśmie naukowym „Information Sciences”.

Wśród badanych dzieł znalazły się m.in.: „1984” George’a Orwella, „Moby Dick” Hermana Melville’a, „Ulysses” Jamesa Joyce’a, „Gulliver’s Travel” Jonathana Swifta, „Gone with the Wind” Margaret Mitchell, „Also sprach Zarathustra” Friedricha Nietzschego, „Der Prozess” Franza Kafki, „Der Zauberberg” Thomasa Manna, „Madame Bovary” Gustave’a Flauberta, „Le Fantôme de L’Opéra” Gastona Leroux, „Il pendolo di Foucault” Umberto Eco, „Giacinta” Luigi’ego Capuany, „Przedwiośnie” Stefana Żeromskiego, „Ziemia obiecana” Władysława Reymonta, „Lalka” Bolesława Prusa, „Anna Karenina” i „Wojna i pokój” Lwa Tołstoja, a także „Bracia Karamazow” Fiodora Dostojewskiego.

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.

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, ale 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 redukowanie interpunkcji na pewno jest działaniem korzystnym i czy czasami nie niszczy wewnętrznej integralności języka?”.

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.

Instytut Fizyki Jądrowej PAN (IFJ PAN) w Krakowie zajmuje się strukturą materii i własnościami oddziaływań fundamentalnych od skali kosmicznej po wnętrza cząstek elementarnych. Wyniki badań – obejmujących fizykę i astrofizykę cząstek, fizykę jądrową i oddziaływań silnych, fazy skondensowanej materii, fizykę medyczną, inżynierię nanomateriałów, geofizykę, biologię radiacyjną i środowiskową, radiochemię, dozymetrię oraz fizykę i ochronę środowiska – są każdego roku przedstawiane w ponad 450 artykułach publikowanych w recenzowanych czasopismach naukowych. Częścią Instytutu jest nowoczesne Centrum Cyklotronowe Bronowice, unikalny w skali europejskiej ośrodek obok badań naukowych zajmujący się terapią protonową nowotworów. IFJ PAN jest członkiem Krakowskiego Konsorcjum Naukowego „Materia-Energia-Przyszłość” o statusie Krajowego Naukowego Ośrodka Wiodącego (KNOW) na lata 2012-2017. Instytut zatrudnia ponad pół tysiąca pracowników. W kategoryzacji MNiSW Instytut został zaliczony do kategorii naukowej A+ w grupie nauk ścisłych i inżynierskich.

KONTAKT:

prof. dr hab. **Stanisław Drożdż**
Instytut Fizyki Jądrowej Polskiej Akademii Nauk
tel. +48 12 6628220
email: stanislaw.drozd@ifj.edu.pl

PRACE NAUKOWE:

„In narrative texts punctuation marks obey the same statistics as words”; A. Kulig, J. Kwapien, T. Stanis, S. Drożdż;
Information Sciences 375 (2017) 98–113; DOI: <http://dx.doi.org/10.1016/j.ins.2016.09.051>

POWIĄZANE STRONY WWW:

<http://www.ifj.edu.pl/>
Strona Instytutu Fizyki Jądrowej Polskiej Akademii Nauk.

<http://press.ifj.edu.pl/>
Serwis prasowy Instytutu Fizyki Jądrowej PAN.

MATERIAŁY GRAFICZNE:

IFJ161130b_fot01s.jpg

HR: http://press.ifj.edu.pl/news/2016/11/30/IFJ161130b_fot01.jpg

Interpunkcja pełni w języku równie ważną rolę co słowa, ujawniają analizy multifraktalne przeprowadzone w Instytucie Fizyki Jądrowej Polskiej Akademii Nauk w Krakowie. (Źródło: IFJ PAN)

IFJ161130b_fot02s.jpg

HR: http://press.ifj.edu.pl/news/2016/11/30/IFJ161130b_fot02.jpg

Prawdopodobieństwo wystąpienia wyrazu (oś pionowa) w zależności od jego rzędu (oś pozioma) w różnorodnych dziełach literatury światowej. Zagadkowe odchylenie od prostej, widoczne na jaśniejszych wykresach dla małych rzędów, znika po uwzględnieniu interpunkcji (wykresy ciemniejsze). (Źródło: IFJ PAN)