

## Jak lingwistyka sądowa przechytryła J.K.Rowling

Autor tekstu: **Virginia Hughes**

Tłumaczenie: **Paulina Wojciechowska**

**N**a początku ubiegłego tygodnia brytyjska gazeta „Sunday Times” wstrząsnęła światem wydawniczym, ujawniając, że Robert Galbraith, świeżo upieczony autor nowej powieści kryminalnej *The Cuckoo's Calling*, to nie kto inny, jak sama J.K.Rowling, autorka bestsellerowej serii książek o Harrym Potterze. Zaraz po tym, w „New York Times” opisano historię, jak redaktorowi do spraw sztuki „Sunday Times”, Richardowi Brooksovi, udało się dokonać tego odkrycia.

Jeden ze współpracowników Brooksa dostał anonimową wiadomość na Twittera. Jej autor twierdził, że Galbraith to tak naprawdę Rowling. Konto „tweetowe” informatora zostało następnie szybko usunięte. Zanim Brooks skonfrontował się z wydawcą *The Cuckoo's Calling*, jego pracownicy poszperali trochę w internecie. Odkryli, że obaj pisarze mieli tego samego wydawcę i agenta. A po konsultacjach z dwójką ekspertów od IT, odkryli, że *The Cuckoo's Calling* i inne książki Rowling wykazują uderzające podobieństwa lingwistyczne. Usatysfakcjonowany tym, że informator tweetowy miał rację, Brooks skontaktował się z Rowling. Wreszcie w sobotę rano, jak donosi „New York Times”, „otrzymał odpowiedź od rzeczniczki Rowling, która powiedziała, że pisarka zdecydowała się „wyznać prawdę”.

Kiedy świat literacki wrzał od pytań, czy anonimowym informatorem nie był w rzeczywistości wydawca Rowling, ja sama chciałam dowiedzieć się, w jaki sposób specje od IT przeprowadzili swoje analizy lingwistyczne. Wczoraj zatelefonowałam do nich i dowiedziałam się nie tylko, jak przeprowadzono „dochodzenie” w sprawie Rowling, ale też dostałam garść informacji o fascynującym świecie lingwistyki sądowej.

Przy pomocy komputerów i wyrafinowanych analiz statystycznych, badacze przeszukują różnego rodzaju słynne teksty, by odkryć wskazówki co do ich autorstwa. Jeszcze bardziej zadziwiający jest fakt, że przeczesują oni teksty nie tylko słynnych autorów, ale blogi, tweety, aktualizacje statusu na Facebooku, a nawet recenzje na Amazonie, by określić styl życia i zwyczaje zakupowe szarego człowieka. Cała sprawa jest paradoksalna. Pisarze starannie dobierają słowa, aby przekazać konkretne treści, ale, jak się okazuje, te same słowa niosą ze sobą dane osobiste, o których nie wiemy, że je przy okazji zdradzamy.

„Z ideą, że szpieg komputerowy może odkryć cechy ukryte w tekście wiąże się coś fascynującego. Takie cechy jak styl pisania, którego czytelnik nie potrafi wykryć, a autor ukryć, to taki odcisk palca lub DNA sposobu, w jaki różni ludzie piszą”, mówi Peter Millican z Uniwersytetu w Oksfordzie, jeden z ekspertów, z którym konsultowała się redakcja „The Sunday Times”.

Cal Flyn, reporterka „The Sunday Times”, wysłał prośbę o pomoc do Millicana i Patricka Juoli, eksperta komputerowego z Duquesne University w Pittsburgu, PA. Flyn przedstawiła im hipotezę, że Galbraith to Rowling i załączyła pięć książek, aby specjaliści mogli tę hipotezę zweryfikować. Wśród tych książek znalazło się oczywiście *Cuckoo*, jak też inna powieść Rowling *The Casual Vacancy*. Pozostałe trzy książki były również brytyjskimi kryminałami, *The St. Zita Society* Ruth Rendell, *The Private Patient* P.D. James oraz *The Wire in the Blood* Vala McDermida.

Juola poddał każdą z książek (a dokładniej, sekwencję kilkudziesięciu tysięcy słów, które składają się na książkę) badaniu przez program komputerowy nazwany JGAAP, nad którym przez ponad dziesięć lat pracował wraz ze swoimi studentami. Porównał *Cuckoo* z pozostałymi książkami stosując cztery różne metody analizy, z których każda koncentrowała się na innym aspekcie analizowanego tekstu.

W jednym z tych tekstów, na przykład, porównywano wszystkie pary słów lub zestawy słów będących obok siebie w każdej z książek. „Metoda ta jest pod wieloma względami lepsza niż porównywanie pojedynczych słów, ponieważ można dzięki niej wychwycić nie tylko to, o czym mówisz, ale także JAK o tym mówisz”, mówi Juola. Np. test ten może pokazać, jakie rzeczy autor opisuje jako „drogie”: drogi samochód, drogie ubrania, drogie jedzenie, itd. „To może być słowo, którego każdy z nas używa, właśnie jak „drogie”, ale w zależności od tego na czym się skupiamy, przekazuje inną myśl.”

Juola przeprowadził także test wyszukujący „znaki n-gramowe”, czyli sekwencje sąsiadujących ze sobą liter. Skupił się na 4 gramach, czyli sekwencjach czteroliterowych. Np., wyszukanie

sekwencji „skok” ujawniało nie tylko użycie tego słowa, ale „skacze”, „skakał”, „skaczący”. „To pozwala nam przyjrzeć się różnym pojęciom i związanym z nimi słowom bez obawy o zastosowany czas i koniugację”.

Takie testy wyszukują względnie rzadko stosowane słowa. Ale również słowa, które najczęściej występują w książce — w języku angielskim takie słowa jak „a, and, of, the” — pozostawiają ukryty odcisk. Dlatego w programie Juoli zestawiono też sto najczęściej występujących słów z każdej książki i porównano niewielkie różnice w częstotliwości ich występowania. W jednej książce „the” mogło zostać użyte w 6%, podczas gdy w innych tylko w 4%.

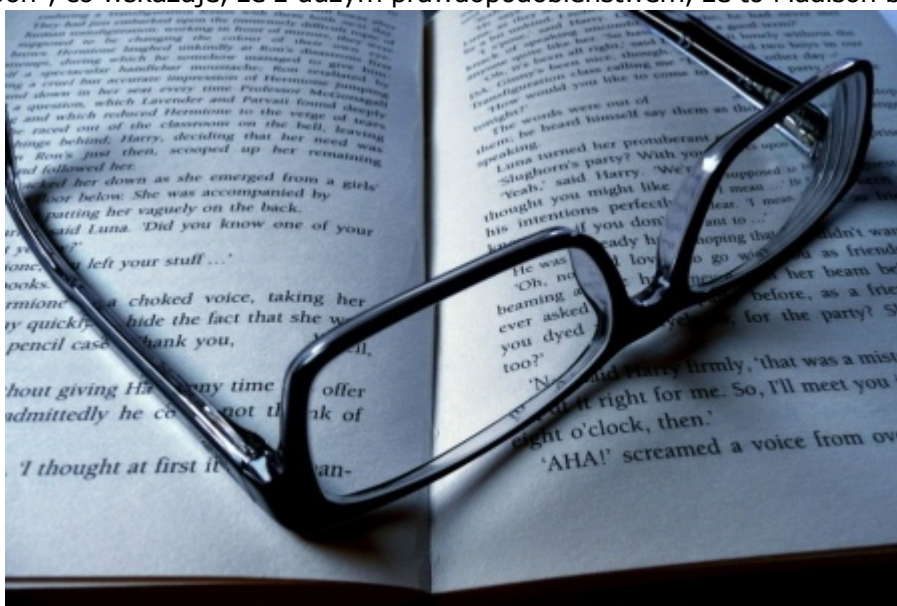
W ostatnim teście Juoli całkowicie odseparowano słowo od jego znaczenia, sortując słowa wyłącznie w zależności od ich długości. Jaki procent książki składa się ze słów trzyliterowych, a jaki z ośmioliterowych? Takie rozkłady są zasadniczo podobne w różnych książkach, ale analizy statystyczne potrafią wykryć pewne subtelne różnice. I ten konkretny test „był bardzo charakterystyczny dla Rowling”, twierdzi Juola. „Długość słów stanowiła jeden z najsilniejszych dowodów na to, że autorem *Cuckoo* jest Rowling.”

Rozwiązanie tych wszystkich zagadek słownych zajęło Juoli około półtorej godziny i wszystkie cztery testy sugerowały, że *Cuckoo* jest bardziej podobna do innej powieści Rowling, *Casual Vacancy*, niż do pozostałych. I to właśnie przekazał Flyn. Nadal jednak nie był całkowicie przekonany co do wyników analizy. Nie mógł przecież wiedzieć, czy prawdziwy autor to nie był inny autor kryminałów, który przypadkiem pisze podobnie do Rowling. „Mógł to być ktoś, kto jest do niej podobny. To samo zdarza się podczas konfrontacji policyjnej”, mówi.

W tym samym czasie za oceanem, Peter Millican przeprowadzał równoległą analizę pisarstwa Rowling. Po otrzymaniu emaila od Flyn, Millican powiedział jej, że potrzebuje więcej danych porównawczych i otrzymał po jednej dodatkowej książce każdego z czterech znanych autorów (skorzystał z powieści *Harry Potter i Insygnia Śmierci* jako drugiej znanej książki Rowling). Podał te osiem książek swojej analizie lingwistycznej przy użyciu programu „Signature”.

W „Signature” stosuje się wyszukaną metodę statystyczną określaną mianem „analizy głównych elementów składowych” w celu porównania wszystkich książek pod względem sześciu cech: długości słowa, długości zdania, długości paragrafu, częstości występowania liter, częstości stosowania różnych znaków interpunkcyjnych i zastosowania określonego słownictwa.

Testy częstości występowania określonych słów mogą być wykonywane na różne sposoby. Jak opisałam wcześniej, Juola przyglądał się parom słów i najczęściej występującym słowom. Inne podejście, które może dać prawie pewną odpowiedź, jest według Millicana porównywanie rzadko stosowanych słów. Klasycznym przykładem są „Federalist Papers”, zbiór esejów autorstwa Aleksandra Hamiltona, Jamesa Madisona i Johna Jaya powstałych podczas tworzenia konstytucji Stanów Zjednoczonych. W 1963 roku badacze zastosowali liczenie słów w celu określenia autorstwa dwunastu z tych esejów, które były napisane albo przez Madisona albo przez Hamiltona. Wykryli, że Madison zwykle używał słowa „whilst” a nigdy „while”, a także „on” raczej niż „upon”. W odróżnieniu od Madisona, Hamilton stosował „while”, nie „whilst”, a „on” i „upon” stosował z tą samą częstotliwością. W dwunastu anonimowych dokumentach nigdy nie użyto słowa „while” i rzadko stosowano „upon”, co wskazuje, że z dużym prawdopodobieństwem, że to Madison był ich autorem.



Millican wykrył kilka potencjalnie wyróżniających się słów podczas „śledztwa Rowling”. Inni autorzy używali słowa „course” (jak w „of course”, - po polsku „oczywiście”), „someone” („ktoś”) i „realized” („zdał sobie sprawę”) trochę częściej niż Rowling. Ale różnica nie była dostatecznie znacząca statystycznie dla Millicana. Dlatego też, podobnie jak Juola, skoncentrował się na najczęściej występujących słowach. Millican wyszukał pięćset najczęściej występujących słów w każdej z książek, a następnie własnoręcznie usunął słowa, które były specyficzne dla tematu, jak „Harry”, „wand” („różdżka”) i „police” („policja”).

Ze wszystkich testów, które Millican może przeprowadzić przy pomocy swojego programu, traktuje te, gdzie porównuje się użycie słów jako najbardziej wiarygodne. „Otrzymujemy wykres, a z wykresu widać jasno, że *Cuckoo Calling* jest ściśle związana z *Harrym Potterem*. Jest także oczywiste, że książki Ruth Rendell i Vala McDermida są do siebie podobne, itd, twierdzi Millican. „Identyfikujemy coś obiektywnego. Trudno opisać, co konkretnie wykrywają testy, ale zdecydowanie wykrywają podobieństwa.” We wszystkich tekstach Millicana, *Cuckoo* wydawała się być najbardziej podobna do znanej powieści Rowling, a ze wszystkich czterech obie książki Rowling były najbardziej podobne do siebie. Millican otrzymał pliki około godziny ósmej w piątek wieczorem. Pięć godzin później napisał do „The Sunday Times”. „Powiedziałem: Jestem prawie pewien, że jeśli to ma być któryś z tych autorów, to jest to Rowling.”

Millican nie po raz pierwszy znalazł się w samym środku debaty dotyczącej książki napisanej przez osobę publiczną. Jesienią 2008 roku, zaledwie na parę tygodni przed wyborami prezydenckimi w USA, otrzymał on email od szwagra republikańskiego kongresmena ze stanu Utah. Ten poinformował Millicana, że wykorzystano jego program „Signature”, który można ściągnąć ze strony naukowca, aby udowodnić, że książka Baracka Obamy, „Dreams from my Father” (*Sny od mojego Ojca*), mogła być napisana przez Billa Ayersa, amerykańskiego terrorystę. „Planowali konferencję prasową w Waszyngtonie, aby zdemaskować Obamę na tydzień przed wyborami i dlatego skontaktowali się ze mną”, wspomina Millican, śmiejąc się. „Znalazłem się w osobliwej sytuacji.” Millican dokonał ponownej analizy i definitywnie wykazał, że *Dreams* nie były napisane przez Ayersa (można o tym przeczytać [tutaj](http://www.philocomp.net/humanities/dreams) (<http://www.philocomp.net/humanities/dreams>)).

Juola też opowiedział mi kilka dziwacznych historii. Kiedyś pracował przy [pewnej sprawie prawnej](http://www.brooklaw.edu/~media/PDF/LawJournals/JLP_PDF/jlp_vol21i2.ashx) ([http://www.brooklaw.edu/~media/PDF/LawJournals/JLP\\_PDF/jlp\\_vol21i2.ashx](http://www.brooklaw.edu/~media/PDF/LawJournals/JLP_PDF/jlp_vol21i2.ashx)) dotyczącej człowieka, który wysyłał anonimowe artykuły do gazet krytykujące rząd pewnego kraju. Groziła mu deportacja z USA, a wiedział, że jeżeli zostanie deportowany, tajna policja tego rządu będzie na niego czekała na lotnisku. Juola udowodnił przy pomocy swoich analiz, że anonimowe artykuły były rzeczywiście napisane przez owego mężczyznę. Dzięki temu mógł on pozostać w USA. „Udało nam się potwierdzić jego autorstwo, co zadowoliło przysięgłych” — mówi Juola. Historia ta, dodaje, pokazuje, jakim potężnym narzędziem może być ten rodzaj nauki. „Istnieje wiele realnych kontrowersji z konsekwencjami dla zamieszanych w nie osób, i to wiele poważniejszych niż sprawa „czy tę dość nijaką powieść napisała słynna autorka”?

Słowa wielu z nas są prawdopodobnie właśnie teraz przez kogoś przeglądane. Niektórzy naukowcy, powiedział mi Juola, pracują nad analizą recenzji na stronach takich, jak Amazon.com. Dzięki tym badaniom można będzie wyeliminować fałszywe pozytywne oceny produktów wystawione przez przedstawicieli firm lub odnaleźć wartościowe wzorce demograficzne.

„Zabawne, ale mogą nam na przykład powiedzieć, że kobiety z zachodnich stanów USA oceniają nasz produkt o półtorej gwiazdki gorzej niż mężczyźni z północnego-wschodu, więc musimy odpowiednio dostosować nasze reklamy. Niewiele firm przyzna się, że przeprowadza takie analizy. Jednak za każdym razem, gdy przeprowadzane jest jakieś badanie, niezależnie od tego, czy jest to śledztwo policyjne, pozwolenie na dostęp do tajemnic państwowych, czy też podanie o pracę, jedną z rzeczy, na które się patrzy, jest czyjś publiczny profil w sieci. Tutaj wszystko jest grą fair.”

[How forensic linguistics outed J.K. Rowling not to mention James Madison, Barack Obama and the rest of us](http://phenomena.nationalgeographic.com/2013/07/19/how-forensic-linguistics-outed-j-k-rowling-not-to-mention-james-madison-barack-obama-and-the-rest-of-us) (<http://phenomena.nationalgeographic.com/2013/07/19/how-forensic-linguistics-outed-j-k-rowling-not-to-mention-james-madison-barack-obama-and-the-rest-of-us>)

Phenomena/National Geographic, 19 lipca 2013r.

### Virginia Hughes

Studiowała neuronauki i dziennikarstwo naukowe, pisuje dla "The New Yorkera", "Popular Science" oraz "Slate", prowadzi blog "Only Human" przy National Geographic Magazine.



[Strona www autora](#)

[Pokaż inne teksty autora](#)

(Publikacja: 27-07-2013)

[Oryginał.](http://www.racjonalista.pl/kk.php/s,9142) (<http://www.racjonalista.pl/kk.php/s,9142>)

Contents Copyright © 2000-2012 Mariusz Agnosiewicz

Programming Copyright © 2001-2012 Michał Przech

Właścicielem portalu Racjonalista.pl jest Fundacja Wolnej Myśli.

Autorem portalu jest Michał Przech, zwany niżej Autorem.

Żadna część niniejszych opracowań nie może być wykorzystywana w celach komercyjnych, bez uprzedniej pisemnej zgody Właściciela, który zastrzega sobie niniejszym wszelkie prawa, przewidziane w przepisach szczególnych, oraz zgodnie z prawem cywilnym i handlowym, w szczególności z tytułu praw autorskich, wynalazczych, znaków towarowych do tego portalu i jakiegokolwiek jego części.

Wszystkie elementy tego portalu, wliczając w to strukturę katalogów, skrypty oraz inne programy komputerowe są administrowane przez Autora. Stanowią one wyłączną własność Właściciela. Właściciel zastrzega sobie prawo do okresowych modyfikacji zawartości tego portalu oraz opisu niniejszych Praw Autorskich bez uprzedniego powiadomienia. Jeżeli nie akceptujesz tej polityki możesz nie odwiedzać tego portalu i nie korzystać z jego zasobów.

Informacje zawarte na tym portalu przeznaczone są do użytku prywatnego osób odwiedzających te strony. Można je pobierać, drukować i przeglądać jedynie w celach informacyjnych, bez czerpania z tego tytułu korzyści finansowych lub pobierania wynagrodzenia w dowolnej formie. Modyfikacja zawartości stron oraz skryptów jest zabroniona. Niniejszym udziela się zgody na swobodne kopiowanie dokumentów portalu Racjonalista.pl tak w formie elektronicznej, jak i drukowanej, w celach innych niż handlowe, z zachowaniem tej informacji.

Plik PDF, który czytasz, może być rozpowszechniany jedynie w formie oryginalnej, w jakiej występuje na portalu. **Plik ten nie może być traktowany jako oficjalna lub oryginalna wersja tekstu, jaki prezentuje.**

Treść tego zapisu stosuje się do wersji zarówno polsko jak i angielskojęzycznych portalu pod domenami Racjonalista.pl, TheRationalist.eu.org oraz Neutrum.eu.org.

Wszelkie pytania prosimy kierować do [redakcja@racjonalista.pl](mailto:redakcja@racjonalista.pl)