Nr rejestracyjny: 2017/26/E/HS2/01019; Kierownik projektu: dr hab. Maciej Jakub Eder

Wielkoskalowa analiza tekstu i metodologiczne podstawy stylistyki komputerowej

W naukach społecznych, naukach przyrodniczych, a nade wszystko w humanistyce jesteśmy świadkami zjawiska określanego mianem „przełom cyfrowy" (*Digital Turn*). Łączy się to z upowszechnieniem technologii cyfrowych, olbrzymią mocą obliczeniową współczesnych komputerów, w szczególności zaś z trudnym do wyobrażenia, wykładniczym przyrostem informacji wytworzonych w przeciągu ostatnich kilku dekad, analizowanych, przeszukiwanych, przechowywanych, a także udostępnionych publicznie w globalnej sieci.

Olbrzymia część danych przechowywanych na serwerach całego świata ma postać tekstu. Wśród najbardziej oczywistych przykładów można wskazać zbiór 30 milionów książek dostępnych w usłudze Google Books, setki hipertekstowych wydań tekstów literackich, tysiące tekstowych stron internetowych, miliony blogów, miliardy "tweetów", etc. „Przełom cyfrowy" oznacza dostęp do niezliczonych danych, ale zarazem stwarza nietrywialne wyzwania. Należą do nich m.in. przeszukiwanie wielkoskalowe, przetwarzanie języka naturalnego, analiza danych, klasyfikacja (bez niej nie byłoby np. automatycznych filtrów spamu), wykrywanie plagiatów, badanie opinii klientów na podstawie wpisów na blogach i wiele, wiele innych. Gwałtowny przyrost materiału sprawia, że pojawiają się zupełnie nowe pytania badawcze oraz – co warto podkreślić – uczeni mogą dziś zweryfikować wiele stawianych przed dziesięcioleciami hipotez. W tzw. humanistyce cyfrowej (*Digital Humanities*) sporo uwagi poświęca się dziś wypracowaniu metod wspomagających ten proces.

Niniejszy projekt, choć dotyczy stylistyki komputerowej tekstów literackich, stanowi niewielki, lecz znaczący wkład w rozwiązanie owego wielkiego wyzwania: jeden z głównych celów projektu zakłada **stworzenie, przetestowanie i zastosowanie innowacyjnej metody porównywania tekstów**, dzięki której będzie możliwe odnajdywanie ukrytych podobieństw stylistycznych i niewidocznych gołym okiem zależności w dużych korpusach tekstowych. Wypracowana metoda stylometryczna (a właściwie zestaw metod) pozwoli klasyfikować teksty pod kątem autorstwa, gatunku, chronologii, treści, typu narracji, płci autora itd. Planujemy jednocześnie, by wypracowane przez nas metody były odporne na zaburzenia w korpusie (np. szum spowodowany dużą liczbą błędów literowych), co jest jednym z ważniejszych, a zarazem niedocenionych wyzwań akwizycji danych internetowych: ogromnych, ale często o bardzo złej jakości. Wypracowaną metodę czeka gruntowne przetestowanie jej skalowalności: chcemy, by testy stylometryczne dało się przeprowadzać na największych nawet korpusach. Metoda taka powinna być również niezależna od badanego języka – będziemy prowadzić badania nad różnymi aspektami stylistycznymi literatury polskiej, angielskiej, łacińskiej i greckiej. Wreszcie, pragniemy wypracować nowatorskie metody przedstawiania relacji tekstowych w formie graficznej (co np. dla kilku tysięcy powieści staje się wyzwaniem niebanalnym).

Projekt jest podzielony na pięć odrębnych, ale bardzo ściśle powiązanych części. Podprojekt **A**, nadrzędny wobec pozostałych części, ma za zadanie wypracowanie nowatorskich metod stylometrycznych; celem podprojektu **B** będzie zastosowanie stylometrii w dziedzinie literatury, na dużych korpusach i w możliwie licznych językach (głównie polskim i angielskim); podprojekt **C**, wykonywany przez doktoranta, ściśle językoznawczy i skierowany na testowanie gramatycznych (składniowych) znaczników stylu; podprojekt **D**, wykonywany przez drugiego doktoranta, stanowiący uszczegółowienie tez stawianych przez pozostałych członków zespołu – będzie to bowiem monografia odmian stylistycznych w obrębie jednej tradycji literackiej, np. studium zmian w literaturze łacińskiej; wreszcie podprojekt **E**, obejmujący rozwijanie metodologii stylometrycznej do badań zmienności wewnątrz (zbiorów) tekstów, np. zmian stylu autora w czasie lub zmienności stylu wewnątrz pojedynczych dzieł.