Teksty Drugie 2007, 3, s.224-228

Magik – nowoczesne narzędzie dla badacza literatury.

Martyna Barczuk, Jerzy Tyszkiewicz

Martyna BARCZUK,
Jerzy TYSZKIEWICZ

Magik – nowoczesne narzędzie dla badacza literatury

W tym artykule przedstawiamy program wymyślony i napisany przez nas z myślą o wykorzystaniu przez środowisko badaczy literatury, a zwłaszcza osoby zajmujące się tekstologią i edytorstwem naukowym. Jest to rodzaj specyficznego edytora tekstu. Jego podstawową funkcją nie jest pisanie i formatowanie, ale porównywanie tekstów napisanych w innych edytorach. Tworzyliśmy go z pomocą prof. Marii Prussak i Marii Kozyry, która była pierwszą użytkowniczką Magika. Mamy nadzieję, że wypróbują go i uznają za pożyteczny także inni badacze. W tej krótkiej nocie chcemy przybliżyć czytelnikom podstawowe zasady działania Magika. Nie jest to instrukcja, jak należy z nim pracować: nie tłumaczymy, gdzie należy kliknąć, żeby uzyskać jakiś konkretny efekt. W istocie program nadal jeszcze zmienia się pod wpływem uwag osób, które go używają, i naszych własnych obserwacji. Pisząc o jego dzisiejszej wersji, utrudnilibyśmy sobie jego ulepszanie, bo zawsze zastanawialibyśmy się, czy jakieś udogodnienie jest na tyle istotne, żeby jego wprowadzenie mogło zrównoważyć niezgodność zachowania programu z opublikowanym już opisem. Naszym zamiarem jest jedynie opis mechanizmu działania programu, który na pewno się nie zmieni, a wyjaśnia wiele jego cech.

Zacznijmy od tego, co nasz program robi. Jego podstawową i prawie jedyną funkcją jest wykrywanie podobieństw pomiędzy dwoma tekstami. Tytułem przykładu, niech pierwszym tekstem będzie cytat: „Hej, odłogiem leży nasza rola, // choć są ziarna, nie ma rąk do siania" z wiersza Jana Kasprowicza *(Poezja 1888)*, a drugim – jego odmiana „Hej, odłogiem leży nasza rola, // choć są ziarna, mało rąk do siania", opublikowana w „Dziennika Poznańskim" w 1886 roku[1].

[1] Przykład podany za: R. Loth *Podstawowe pojęcia i problemy tekstologii i edytorstwa naukowego*, Wydawnictwo IBL PAN, Warszawa 2006, s. 137.

Umieśćmy oba fragmenty jeden pod drugim, ustawiając pod sobą kolejne jednostki tekstu (za jednostki uważamy: słowa, znaki przestankowe i odstępy rozdzielające je od siebie):

> Hej, odłogiem leży nasza rola, choć są ziarna, nie ma rąk do siania,
> Hej, odłogiem leży nasza rola, choć są ziarna, mało rąk do siania.

Zauważamy 23 pary jednostek zgodne ze sobą w obu wersjach: 8 par słów, 3 pary przecinków i 12 zgodnych par odstępów rozdzielających słowa, ponadto 5 par jednostek niezgodnych (w tym jedną, w której słowo „siania" jest połączone w parę z kropką) i jeden brak pary dla przecinka na końcu. Oczywiście, narzuca się sposób, jak można to zrobić lepiej, wprowadzając do drugiego tekstu specjalne „puste słowo" oznaczone „–".Wtedy możemy otrzymać porównanie

> Hej, odłogiem leży nasza rola, choć są ziarna, nie ma rąk do siania,
> Hej, odłogiem leży nasza rola, choć są ziarna, mało – rąk do siania.

Tym razem znajdujemy 26 par jednostek zgodnych ze sobą w obu wersjach: 11 złożonych ze słów, 3 ze znaków przestankowych i 12 z odstępów. Mamy też 2 pary niezgodne i jedną parę zawierającą puste słowo.

Nasz program wszystkim możliwym porównaniom tej postaci przypisuje liczbę punktów, która jest wynikiem zsumowania punktów przyznawanych za każdą parę jednostek oraz odjęcia punktów ujemnych za każdą parę niezgodną, każdy brak pary lub użycie pustego słowa. Wartość punktowa pojedynczej pary każdego z tych rodzajów została przez nas dobrana na podstawie eksperymentów i jest niezmienna. Żeby wytłumaczyć działanie Magika umówmy się, że za zgodne pary damy po 5 punktów, za niezgodne minus 3 punkty, zaś za użycie pustego słowa lub brak pary minus 5 punktów (prawdziwe wartości stosowane w programie są ułamkami i utrudniłyby niepotrzebnie rachunki).

W tej sytuacji pierwsze porównanie daje $23\times5-5\times3-1\times5=95$ punktów, podczas gdy drugie $26\times5-2\times3-1\times5=199$ punktów. Drugie porównanie okazuje się zatem lepsze, oczywiście zgodnie z naszymi odczuciami. Mechanizm działania stosowanego przez nas algorytmu sprowadza się do czysto mechanicznego wyszukania takiego porównania, które daje najwyższą możliwą liczbę punktów. W rzeczywistości program nie rozpatruje wszystkich możliwych porównań, bo jest ich dosłownie nieskończenie wiele: dla tekstów o długości już od około 150 jednostek więcej niż atomów w obserwowalnej części Wszechświata. Mimo to porównanie, które ma najwyższą z wszystkich punktację, zostaje na pewno znalezione. Następnie na ekranie program wskazuje w tym właśnie porównaniu (poprzez odpowiednie pokolorowanie; poniżej użyliśmy podkreślenia i podkreślenia na szarym tle do reprezentowania dwóch różnych kolorów w czarno-białym druku) fragmenty tekstu drugiego, których brak w tekście pierwszym. Puste słowa są reprezentowane przez oznaczenie [–]. W naszym przykładzie wyglądałoby to tak:

> Hej, odłogiem leży nasza rola, choć są ziarna, nie ma rąk do siania,
> Hej, odłogiem leży nasza rola, choć są ziarna, mało [–] rąk do siania.

Nasz program umożliwia badanie wielu wersji tego samego tekstu, przy czym są one porównywane do siebie kolejno: pierwsza z drugą, druga z trzecią itd. W ten sposób, przy odpowiednim doborze kolejności wersji, w każdej z nich widzimy zaznaczone te fragmenty, które po raz pierwszy pojawiają się właśnie w niej. Kolory z wersji poprzednich zostają zachowane, dzięki czemu w ostatniej wersji możemy śledzić całą ewolucję tekstu. Widać to na przykładzie krótkiego fragmentu *Ody do młodości* Adama Mickiewicza, wers 21, kolejno w wersjach: „Kopia Chelmickiego", „Autografia" i tekst jednolity[2]:

Sam się żaglem, sternikiem, okrętem,
Sam sobie sterem, żeglarzem, okrętem,
Sam sobie sterem, żeglarzem, okrętem;

Przy tej kolejności porównywanych tekstów, podkreślone fragmenty to te, które pojawiają się w drugiej wersji, a podkreślony średnik na szarym tle po raz pierwszy zastępuje przecinek w wersji trzeciej. Jest to w istocie rzeczy uzyskane automatycznie kolacjonowanie[3].

Utworzone za pomocą Magika porównania tekstów mogą zostać w nim opatrzone przypisami i zachowane w takiej postaci na dysku. W zasadzie umożliwia to stworzenie elektronicznej wersji pełnej, krytycznej edycji dzieła, z podaniem wszystkich jego wersji *in extenso* i wskazaniem wszelkich, nawet najdrobniejszych różnic pomiędzy nimi. Wobec znanych trudności z drukiem obszernych, krytycznych wydań literatury pięknej, pewnym rozwiązaniem może być druk tylko tekstu jednolitego, z przeniesieniem całego aparatu krytycznego do dołączanej do książki na płycie lub udostępnianej w internecie wersji elektronicznej, tworzonej za pomocą Magika i przeznaczonej do odczytywania za jego pomocą. Z myślą o takim wykorzystaniu naszego programu będzie on rozprowadzany na zasadach licencji *Gnu General Public Licence (GPL)*. Oznaczać to będzie przekazanie użytkownikom praw do uruchamiania programu w dowolnym celu, analizowania sposobu jego działania i dostosowywania go do swoich potrzeb, kopiowania oraz udoskonalania, a także publikowania własnych poprawek i ich kodu źródłowego.

Ubocznym efektem mechanizmu działania naszego programu jest to, że można go też użyć w zupełnie inny, niespodziewany sposób: do wyszukiwania w długich tekstach cytatów, których brzmienia dokładnie nie pamiętamy. Wystarczy w tym celu porównać ze sobą zapisany w pliku cytat z treścią całego dzieła.

Po przedstawieniu zalet Magika, wypada wspomnieć o jego ograniczeniach. Jak łatwo się domyślić, nawet niewielka zmiana w jednym z porównywanych fragmentów może spowodować istotne zmiany w wyniku działania programu. Gdybyśmy zrobili literówkę w wierszu Kasprowicza, wstawiając spację do wnętrza słowa „mało", nowe najlepsze porównanie byłoby takie:

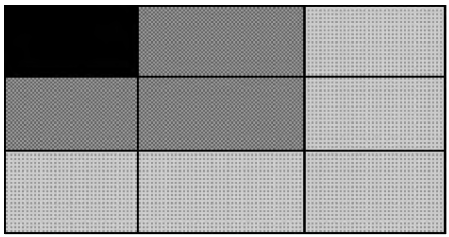
2 A. Mickiewicz *Dzieła wszystkie*, t. 1: *Wiersze*, Ossolineum, Wrocław 1971, s. 283.

3 R. Loth *Podstawowe pojęcia...*, s. 89-91 i ilustracja przed s. 89.

> Hej, odłogiem leży nasza rola, choć są ziarna, nie ma – rąk do siania,
> Hej, odłogiem leży nasza rola, choć są ziarna, [–] ma to rąk do siania.

Stąd nasza decyzja, żeby bardzo ograniczyć możliwość edycji tekstów w Magiku, bo każda ich zmiana powoduje, że dotychczasowe porównanie przestaje być wiarygodne i w zasadzie wymaga ponownego przeprowadzenia całego czasochłonnego obliczenia. Zresztą Word czy OpenOffice dają tak wielkie możliwości edycyjne, że nie miałaby sensu próba ich naśladowania. Postanowiliśmy, że z wyjątkiem najniezbędniejszych sytuacji do edycji używa się innych programów, a Magik służy tylko do porównywania.

Każdy użytkownik naszego programu zauważy, że wyliczanie porównań długich tekstów jest czasochłonne. Istotną informacją jest to, że nakład pracy wykonywanej przez komputer przy obliczaniu najlepszego porównania jest w przybliżeniu proporcjonalny do i l o c z y n u długości obu tekstów. Oznacza to, że dobrym modelem tej ilości pracy jest pole prostokąta o bokach równych co do długości odpowiednio liczbom słów w obu tekstach. Patrząc na poniższy obrazek, bez trudu zauważymy, że jeśli na przykład podwoimy długości obu porównywanych fragmentów,



powinniśmy spodziewać się, że czas pracy programu wzrośnie czterokrotnie, a przy potrojeniu długości wzrost czasu pracy będzie dziewięciokrotny. Można tej reguły używać do szacowania czasu niezbędnego do przeprowadzenia całego porównania za pomocą próby wykonywanej na fragmentach.

Na zakończenie ciekawostka: Magik wykorzystuje do wyszukiwania najlepszych porównań algorytm Hirschberga, który jest adaptacją znanego w informatyce algorytmu obliczania tak zwanej „odległości edycyjnej". Podobnymi modyfikacjami są też algorytmy wykorzystywane w biologii do badania stopnia podobieństwa kodów DNA różnych organizmów. Tak więc badania Marii Kozyry nad genezą dramatu *Orfeusz* Anny Świrszczyńskiej i badania nad genomem człowieka nie są wcale tak odległe od siebie, jak by się mogło wydawać na pierwszy rzut oka.

Abstract

Martyna BARCZUK, Jerzy TYSZKIEWICZ
Warsaw University

Magician: a scholar's modern tool

This article presents our programme written in view of it being taken advantage of by the literary scholars' milieu, particularly, those dealing with textology and scientific editorship. The programme enables clear presentation of differences between various versions of a text. We try to describe to the reader a general idea of how the program may operate and to make them aware of the barriers limiting the programme.