# Zastosowanie archiwaliów i nowoczesnych technologii w służbie badania języka

Agnieszka Jaworska
Daniel Dzienisiewicz
Rafał Jaworski

Jak zauważa profesor Jerzy Topolski, „źródła nie mogą odbijać przeszłości”, jednakże pozwalają one na zrozumienie i poznanie zdarzeń i rzeczy minionych (Topolski, 2001). Należy tutaj zastanowić się, czym w takim wypadku jest źródło oraz w jaki sposób może ono posłużyć w badaniu przeszłości. Podstawowym źródłem w badaniach archiwistycznych jest zachowana dokumentacja. Dzięki jej różnorodności, można przeprowadzać badania na wielu płaszczyznach.

Jednym z zastosowań archiwaliów jako źródła jest wykorzystanie ich w badaniach lingwistycznych. Zachowana dokumentacja pozwala zaobserwować, w jaki sposób zmieniał się język. Większość z tego typu źródeł reprezentuje język formalny (zaliczyć możemy do nich np. protokoły), jednakże odnaleźć można również język specjalistyczny (w szczególności w np. dokumentacji technicznej) oraz potoczny (występujący np. w zeznaniach zawartych w aktach sądowych). Należy zwrócić uwagę, że dla szerszego pogłębienia poruszanej problematyki niezbędne jest przeprowadzenie badań ilościowych i jakościowych, które są możliwe dzięki procesowi digitalizacji zachowanej dokumentacji.

Digitalizacja archiwaliów przeprowadzana jest w wielu archiwach polskich i zagranicznych. Archiwa Stanów Zjednoczonych szczycą się udanym zdigitalizowaniem swoich zasobów. Każdy proces digitalizacji archiwów jest jednak niezwykle kosztowny i czasochłonny. Wymaga między innymi dostępności specjalistycznych urządzeń skanujących, a przede wszystkim angażuje do pracy wiele osób. Nie są bowiem dostępne urządzenia, które w pełni zautomatyzowany sposób są zdolne skanować dokumenty w różnych formatach lub zorganizowane w różny sposób (zawarte w teczce, spięte zszywkami lub w formie książki, czy zeszytu). Co więcej, bezpośredni nadzór człowieka nad procesem skanowania jest konieczny także ze względu na zapewnienie należytej ochrony przed przypadkowym zniszczeniem cennych zbiorów archiwalnych.

Na samym skanowaniu nie kończy się jednak proces digitalizacji dokumentów archiwalnych. Kolejnym krokiem jest zapewnienie przestrzeni dyskowej na przechowywanie wykonanych zdjęć dokumentów archiwalnych. Nowoczesne skanery i aparaty cyfrowe charakteryzują się znacznej wielkości matrycą, co pozwala na wykonywanie zdjęć w wysokiej rozdzielczości, często niezbędnej do prawidłowego odczytania dokumentów. Takie zdjęcia wymagają znacznej przestrzeni dyskowej, a co za tym idzie, koszt ich przechowywania jest wysoki.

Powyższe czynniki powodują, że zdigitalizowanie wszystkich zbiorów archiwalnych zgromadzonych w polskich archiwach (celem np. opublikowania ich za pośrednictwem Internetu), nawet przy obecnym stanie zaawansowania technologicznego, jest niezwykle trudne. Możliwa jest natomiast digitalizacja wybranych archiwaliów z różnych okresów. Dane pozyskane w ten sposób doskonale nadają się do przeprowadzenia zarówno ilościowej, jak i jakościowej analizy językoznawczej.

Żeby jednak była ona możliwa, konieczne jest wykonanie jeszcze jednej operacji na zdjęciach archiwaliów. Jest nią proces ekstrakcji tekstu ze zdjęcia dokumentu, nazywany OCR (Optical Character Reading). W odróżnieniu od skanowania, OCR jest wykonywany całkowicie automatycznie przy pomocy komputera, wyposażonego w odpowiednie oprogramowanie (np. ABBYY Fine Reader lub Tesseract). Choć wyniki operacji ekstrakcji tekstu są zazwyczaj zadowalające, należy pamiętać o najważniejszym ograniczeniu technologicznym tej metody: możliwości odczytywania tekstu jedynie z druków lub maszynopisów (nigdy z pisma odręcznego). Gdy jednak uda się przeprowadzić OCR na dokumentach archiwalnych, badacze języka otrzymują w formie tekstowej znacznej wielkości zbiór tekstów w języku polskim. Co więcej, w przypadku każdego tekstu dysponują informacją na temat roku (często również miesiąca i dnia) napisania

tekstu, kategorii tematycznej czy miejsca wydania.

Przykładem wniosku, który został wyciągnięty podczas lingwistycznych dociekań prowadzonych na zdigitalizowanym materiale tekstowym, jest przede wszystkim jego nieoceniona przydatność w badaniach nad (re)datacją słownictwa XX wieku (por. np. Wierzchoń 2008, 2009). Dzięki prężnie rozwijającej się digitalizacji archiwaliów oraz bibliotekom cyfrowym i publikowanym przez nie w Internecie masom tekstowym, proces datowania słownictwa stał się wydajniejszy niż kiedykolwiek przedtem. Nielimitowany dostęp do zgromadzonych materiałów tekstowych otwiera bowiem przed badaczem możliwość szybkiego uchwycenia najwcześniejszego wystąpienia poszukiwanej jednostki języka (leksemu, reproduktu) względem danego zbioru tekstów, przy czym im większy i bardziej reprezentatywny jest ów zbiór, tym większe prawdopodobieństwo odnalezienia użycia obiektywnie najwcześniejszego. Dokumenty archiwalne, na których przeprowadzony został OCR, pozwalają na wyszukiwanie form wyrazów i fraz bez konieczności każdorazowego czytania dokumentów, co znacznie skraca czas dotarcia do badanych bytów językowych. Szczegółowa informacja określająca rok (miesiąc, dzień) wydania sprawia, iż lingwista dysponuje wszelkimi danymi potrzebnymi do poprawnej datacji jednostki. Wydajna metoda chronologizacji słownictwa pomaga ustalić, kiedy poszczególne wyrazy pojawiły się w polszczyźnie i tym samym wnikliwie opisać całe jej bogactwo i różnorodność.

Lingwochronologizacja wspomóc może także m.in. słowotwórstwo gniazdowe poprzez ustalenie czasu pojawienia się derywatów w języku i ukazanie ich w aspekcie diachronicznym w odróżnieniu od dotychczasowych opisów synchronicznych obecnych w słownikach słowotwórczych (por. Wierzchoń 2010). W szerszej perspektywie możliwe jest także ustalenie czasu, który minął np. od wynalezienia jakiegoś urządzenia (np. telefonu) i werbalizacji jego nazwy w tekstach, a także ustalenie obecności wariantów występujących przed ukonstytuowaniem się powszechnie używanego terminu.

Innym przykładem wykorzystania archiwaliów w badaniach językowych jest datowanie słownictwa według ścisłego kryterium tematycznego, co umożliwia sporządzenie szczegółowego opisu rozwoju danej dziedziny, np. poprzez prześledzenie rozwoju terminologii medycznej lub sakralnej. Ponadto, dzięki zgromadzonym zbiorom tekstów, możliwe jest także ustalenie wzrostu produktywności danego wyrazu w tekstach oraz udzielenie szerszego komentarza, pod wpływem jakich czynników lub w stosunku do jakich podmiotów był on używany. Historyczny przekrój tekstów pozwala także na przeprowadzenie wiwisekcji semantyki poszczególnych jednostek i ustalenie ich znaczenia w różnych przedziałach czasowych.

Powyższe badania nie byłyby jednak możliwe bez ogromnego nakładu pracy ze strony archiwistów. Odpowiednio zachowana dokumentacja może być źródłem wiedzy o języku, jego fluktuacjach oraz zastosowaniach. Podczas wystąpienia zaprezentowane zostaną dotychczasowe osiągnięcia poznańskich badaczy w zakresie pracy z archiwalnymi materiałami udostępnianymi przez biblioteki cyfrowe, a także przedsięwzięcia planowane oraz te, nad którymi trwają prace w chwili obecnej. Ponadto w możliwie szczegółowy sposób przedstawione zostaną metody pracy ze zdigitalizowanym materiałem tekstowym.