Marek A. Iwanowski

Ku humanistyce cyfrowej – filolog pyta o bibliografię zawartości Internetu

Recenzent:
Jan Pudzikowski

Wydawca:

Mila Hoshi

Warszawa 2014

ISBN 978-83-939543-1-5

Nakł. 5 egz. papierowych + egz. elektroniczny
(dostępny w Internecie)

Lubelskim mistrzom humanistyki cyfrowej

– Co jest zawartością Internetu dla filologa? Filolog interesuje się tekstami.

Przytoczone pytanie i próba odpowiedzi na nie, zaczerpnięte z ulotnej rozmowy środowiskowej, niech będą punktem wyjścia dla niniejszej pracy. Zarysowane stanowisko jest nieco zbyt ostre, gdyż filolog oprócz tekstów dostrzega w Internecie – a faktycznie na stronach internetowych wyświetlanych na ekranie lokalnego komputera – również obrazy, ilustracje, filmy, wykresy, tabele, przyciski oraz wiele innych obiektów, często interaktywnych; jednak w rzeczy samej dla filologa najważniejsze są teksty.

Porządnych bibliografii zawartości Internetu przeznaczonych dla filologów jeszcze nie ma. Nie utarła się nawet nazwa dla takiego narzędzia. A warto by było pomyśleć, gdyż termin bibliografia mocno kojarzy się z tradycyjnymi bibliotekami, a mało z Internetem. Zapewne wkrótce ktoś zaproponuje ładną i adekwatną etykietkę, niemniej teraz na roboczo można by rozważyć przyjęcie jednej z następujących nowych nazw: stronografia internetowa, WWW-grafia, w3grafia, trzywugrafia. Przed przejściem do sprawy zasadniczej przedstawiam bliskie otoczenie tematu, gdyż warto by powiększała się liczba filologów w nim zorientowanych.

Czytając tekst, filolog, a w szczególności językoznawca[1], np. gramatyk czy też tłumacz, interpretuje go zgodnie z własną siatką pojęć i reguł interpretacji. Oprócz pojęć i reguł ściśle językoznawczych filolog stosuje również pojęcia i reguły odnoszące się bezpośrednio do tzw. rzeczywistości pozajęzykowej, czyli codziennego doświadczenia zawodowego i życiowego. Jeśli studiując dany tekst, filolog natrafi na użycie nieznanej mu jednostki języka i nie może jednoznacznie określić jej znaczenia, wtedy sięga do źródeł, w których spodziewa się znaleźć stosowny opis, definicję, gdzie oprócz tekstu może natrafić na obrazy, wykresy, tabele. W trudniejszych przypadkach powinien porozumieć się ze specjalistami z danej dziedziny lub nawet jako obserwator, dbając by nie wyrządzić sobie szkody, doświadczyć danego obiektu czy sytuacji w sposób jak najmniej wpływający na istotne cechy tego, co postanowił poznać przez obserwację w tzw. rzeczywistości pozajęzykowej.

Kiedy filolog pracuje, wtedy zgodnie z celem pracy przekształca badany tekst wejściowy na tekst wynikowy. Badany tekst wejściowy nie może być ciągiem przypadkowych znaków, musi on spełniać szereg warunków, które w wielkim uproszczeniu można przedstawić w ten sposób: badany tekst wejściowy składa się z ciągu w miarę sensownych zdań. Gromadząc i selekcjonując teksty

[1] Pozostaję przy bardzo ogólnym określeniu pojęcia filolog, choć gdy mowa o konkretnych czynnościach, to zręczniej było by mówić o specjaliście konkretnej dziedziny.

do badań, filolog zależnie od potrzeb zwraca uwagę na ich cechy metatekstowe takie jak: autor, tytuł, czas powstania, gatunek i rodzaj literacki.

Użytkownik Internetu zwykle korzysta na lokalnym komputerze z okienkowego systemu operacyjnego Windows, Linux lub Mac, z którym współdziała przeglądarka stron internetowych taka jak Internet Explorer lub Mozilla Firefox. Jeśli strona internetowa nie mieści się na ekranie komputera, to można ją przeglądać za pomocą przewijania w pionie lub poziomie. W zasadzie strony internetowe nie podlegają edycji przez odbiorcę (czytelnika), z wyjątkiem np. pól przeznaczonych do wpisywania loginów, haseł czy pól edycyjnych programów poczty elektronicznej.

Dostawca (autor) strony internetowej najczęściej dostarcza zestaw stron – zwany witryną lub lokalizacją – wśród których jedna pełni rolę strony głównej, a reszta to jej podstrony. Aby przejść z danej strony do jej podstrony, dana strona musi zawierać tzw. link (łącze), który jest adresem podstrony. Link często jest ukryty pod etykietą lub obrazem przycisku, które są obiektami reagującymi na kliknięcie powodujące przejście do podstrony. Zwykle etykieta oraz przycisk są zaopatrzone w tekst służący jako drogowskaz. Powiązania między stronami muszą umożliwiać dotarcie ze strony głównej do dowolnej podstrony na danej witrynie.

Klasycznym sposobem dotarcia do konkretnej podstrony jest wpisanie adresu internetowego strony głównej danej witryny do pola adresu przeglądarki internetowej, a następnie nawigowanie zgodnie z linkami od strony głównej przez kolejne podstrony aż do strony docelowej. Zwykle struktura powiązań podstron witryny jest drzewiasta, ale trafiają się również cykle. Jeśli odesłanie zamykające cykl prowadzi do strony głównej, to można taką sytuację uznać za jeszcze prawidłową, gdyż służy do szybkiego powrotu do początku przeglądania witryny. Duże wątpliwości budzą natomiast odesłania między podstronami zamykające się w cykle łączące podstrony należące do różnych gałęzi, gdyż mogą powodować uciążliwe błądzenie użytkownika.[2] Zdarzają się nawet odesłania do innych witryn. Jeśli odesłania nie są dobrze opisane, to można przeoczyć przejście na inną witrynę.

Jeśli filolog trafi na interesujące go użycie jednostki językowej, to powinien udokumentować znalezisko, zapamiętując je na dysku twardym lokalnego komputera. Dla zestawu składającego się z systemu operacyjnego Windows XP, przeglądarki Internet Eksplorer 8, edytora Word 2003, programu graficznego Paint procedura wygląda następująco:

[2] Im więcej takich cykli, tym mniej można mówić o strukturze drzewiastej, a bardziej po prostu o sieci.

1. zrobić zrzut ekranu;[3]
2. skopiować do pliku tekstowego pełny adres oglądanej strony internetowej;[4]
3. zapamiętać na dysku twardym oglądaną stronę internetową;[5]
4. skopiować do pliku tekstowego (o rozszerzeniu TXT, RTF, DOC itp.) użycie jednostki językowej wraz z kontekstem;[6]
5. zapamiętać w pliku tekstowym datę wpisu na stronie (lub datę powstania strony) oraz aktualną datę.

Przeprowadzenie wymienionych czynności wymaga nieco więcej umiejętności niż podstawowa obsługa powszechnie znanego programu edycyjnego Word firmy MicroSoft czy jego darmowego odpowiednika Writer organizacji OpenOffice. Zapamiętanie tekstu w postaci kodów znaków oraz w postaci obrazu pozwala na pełne udokumentowanie wystąpienia danej jednostki językowej. Obraz pozwala na późniejszą fotodokumentację znaleziska, a tekst w postaci kodów znaków może służyć celom wyszukiwawczym, gdy znalezisk nazbiera się tak dużo, że warto powiązać je w bazę danych.

W zasadzie każde kopiowanie tekstu w postaci kodów znaków, które wykorzystuje wewnętrzne mechanizmy komputera służące do kopiowania ciągów kodów znaków, może powodować błędy, gdy nie zgadzają się strony kodowe komputerów, programów, okienek lub innych obiektów albo strona internetowa od strony technicznej została napisana niestarannie. Każde przejście do oglądania innej witryny, a nawet innej strony internetowej, może powodować wspomniane niezgodności. Przykładem kopiowania, które może prowadzić do błędów z powodu niezgodności stron kodowych, jest, wydawałoby się proste i często stosowane, kopiowanie za pomocą skrótów klawiaturowych [Ctrl + C] i [Ctrl + V], czyli kopiowanie z użyciem schowka systemu Windows. Najłatwiej przeoczyć błędy kopiowania, gdy dotyczą one niewielkiej liczby znaków, np. pojedynczego znaku diakrytycznego. Zatem wynik kopiowania za pomocą skrótów klawiaturowych należy sprawdzać czytając skopiowany tekst.

W zasadzie nie ma ograniczenia liczby podstron wchodzących w skład danej witryny. Może ich być wiele tysięcy. Wtedy z pomocą może przyjść lokalna

[3] Zrzut ekranu to obraz ekranu komputera zapamiętany na twardym dysku jako plik graficzny. W systemie Windows zrzut ekranu przesyłany jest do schowka z pomocą skrótu klawiaturowego [Shift + Print Scrn (Screen)], a następnie obraz przesyłany jest do programu graficznego (np. Paint, IrfanView) za pomocą skrótu klawiaturowego [Ctrl + V], gdzie V oznacza napis na klawiszu, a nie dużą literę pisaną z użyciem klawisza Shift.

[4] Zaznaczenie aktualnego adresu wyświetlanego w oknie przeglądarki, następnie skopiowanie go do schowka za pomocą skrótu klawiaturowego [Ctrl + C], po czym uaktywnienie okna edytora pokazującego stosowny plik tekstowy i wklejenie zawartości schowka za pomocą skrótu [Ctrl + V].

[5] Użycie skrótu klawiaturowego [Ctrl + S] lub opcji Plik | Zapisz jako. Ta operacja nie zawsze da się wykonać za pomocą Internet Explorera. Natomiast przeglądarka FireFox lepiej sobie radzi, ale częściej gubi jakieś składniki zapamiętywanej strony.

[6] Zaznaczenie kursorem myszy tekstu, następnie skopiowanie go do schowka za pomocą skrótu klawiaturowego [Ctrl + C], po czym uaktywnienie okna edytora pokazującego stosowny plik tekstowy i wklejenie zawartości schowka za pomocą skrótu [Ctrl + V]. Może się zdarzyć, że niektóre okienka mają wyłączoną możliwość zaznaczania tekstu oraz kopiowania do schowka.

wyszukiwarka umieszczona na którejś ze stron witryny, zwykle na stronie głównej. Lokalna wyszukiwarka witryny działa tak, jak zaprojektował ją programista, a więc dopóki użytkownik jej nie wypróbuje, to się nie dowie, jakiego rodzaju zapytania są skuteczne i jakie wyniki są zwracane. Używając jedynie przeglądarki, raczej trudno zorientować się co tak naprawdę jest przeglądane, gdyż może być przeglądana baza danych podłączona do witryny, albo zestaw stron witryny, albo pliki w katalogach, do których nie można dojść metodą nawigowania po podstronach witryny.

Lokalnie na stronie internetowej można próbować szukać tekstu za pomocą opcji *Edycja | Znajdź na tej stronie* lub za pomocą skrótu klawiaturowego [Ctrl + F] z jednoczesnym wpisaniem wzorca wyszukiwawczego.

Niektóre witryny składają się z wielu milionów stron (podstron) internetowych, dobrym przykładem może być anglojęzyczna witryna MSDN firmy MicroSoft zawierająca dokumentację systemu operacyjnego Windows, programów pakietu MicroSoft Office oraz stowarzyszonych języków jak MS C++ czy Visual BASIC. Z witryn polskojęzycznych bogate bywają internetowe fora dyskusyjne zawierające setki tysięcy wypowiedzi dyskutantów, jak np. witryna *forum.wiara.pl*, która zawierała 770441 wypowiedzi, gdy było pisane to zdanie.

Na stronie internetowej może być zagnieżdżony jeden lub więcej dokumentów o formatach takich jak np. PDF (*Portable Document Format*), DjVu (*déjà vu*), XML (*Extensible Markup Language*), DOC (dokument edytora Word) oraz TXT (czysty plik tekstowy np. wytworzony za pomocą programu Notatnik). Duże zasoby dokumentów o podanych wyżej formatach można znaleźć na witrynach instytucji lub organizacji takich jak Sejm, Senat, ministerstwa, Urząd Patentowy, urzędy samorządowe, biblioteki elektroniczne.

Największym kłopotem dla filologa poszukującego materiałów źródłowych do badań jest nawigacja po stronach witryn internetowych. Trzeba z góry znać adres witryny, wiedzieć co tam można znaleźć; nawigować po stronach witryny za każdy razem dostosowując się do konkretnego rozwiązania, gdzie układ składników na stronach bywa bardzo różny i różne są powiązania między stronami. Trudno jest zapamiętać sposoby dojścia do danej strony, mało kto dokumentuje to, jak do niej doszedł, a potem trudno jest mu trafić w to samo miejsce.

W e-bibliotece trzeba przejść przez katalog, często wspólny dla wydawnictw papierowych. Najczęściej trzeba znać autora lub tytuł. Jeśli jest możliwość tzw. wyszukiwania złożonego, to zwykle można wyszukiwać według dowolnych słów z rekordu bibliotecznego połączonych operatorami logicznymi. Czasem można użyć słów kluczowych. Natomiast e-katalogi rzeczowe należą do rzadkości, co jest zrozumiałe, gdyż z natury rzeczy mają nieco bardziej skomplikowaną strukturę i trudniej jest je zaimplementować.

Mimo tych udogodnień filolog językoznawca nie może czuć się komfortowo, gdyż mało kiedy interesuje się jednym dziełem np. konkretnego autora. Raczej wolałby wyselekcjonować grupę dzieł konkretnego gatunku czy rodzaju, speł-

niającą jakieś dodatkowe kryterium, a następnie przeglądać je strona po stronie lub wyrywkowo, zależnie od przyjętej metody. Natomiast danymi bibliograficznymi dzieła i lokalizacją w dziele interesuje się dopiero po znalezieniu miejsca wystąpienia jednostki językowej. Warto zauważyć, że często nie szuka konkretnej jednostki, gdyż zna jedynie jej ogólniejszą charakterystykę, np. może poszukiwać nowych słów.

W momencie pytania o dane bibliograficzne dzieła, gdy znalezione zostało miejsce wystąpienia jednostki językowej (np. w dużym korpusie), a o dziele wiadomo jedynie, że należy do dużego zbioru dzieł źródłowych, następuje jak gdyby odwrócenie perspektywy, gdyż z poziomu tekstu zadawane jest pytanie o dane bibliograficzne. Tego rodzaju spojrzenie jest bliskie Janowi Wawrzyńczykowi, który rozwija zupełnie nowy nurt bibliograficznego wspomagania badań językoznawczych, a szczególnie leksykograficznych (Wawrzyńczyk 1998, Wawrzyńczyk 2000-2012). Należy być świadomym, że dla prac leksykograficznych prowadzonych z tej perspektywy fundamentalne znaczenie mają opracowania językoznawcze na temat danej jednostki językowej oraz że status tych opracowań jest inny niż status materiałów źródłowych. Zatem należy zwracać uwagę na skutki umieszczania w korpusach zarówno materiałów źródłowych, jak i ich opracowań językoznawczych.

Gdy oglądamy dokument na ekranie komputera, wtedy widok ilustracji pochodzi z pliku graficznego, a ten z kolei pochodzi ze skanu, zdjęcia itp. Natomiast widok litery (czcionki) tekstu poza ilustracją jest obliczany z opisu czcionki zawartego w pliku z fontami, a podstawą wyliczenia jest kod czcionki. Stąd można mówić o tekście obliczonym (złożonym z czcionek), który charakteryzuje się tym, że wszystkie widoki (realizacje) tej samej litery są identyczne, gdy zachowane są stosowne parametry jak stopień pisma, krój itp. Jeśli zauważymy, że na ilustracji może pojawić się obraz tekstu, a ilustracja jest częścią lub całością obrazu pochodzącego ze skanowania lub zdjęcia fotograficznego, to taki tekst można nazwać tekstem graficznym, który charakteryzuje się tym, że widoki różnych realizacji tej samej litery są różne. Gdy w skrajnym przypadku cała strona jest ilustracją (obrazem pochodzącym ze skanu, zdjęcia itp.), wtedy cały tekst znajdujący się na niej jest tekstem graficznym.

Zespoły złożone z informatyków programistów oraz językoznawców piszą programy automatycznego przetwarzania tekstów naśladujące niektóre aktywności językoznawcze jak analiza morfologiczna oraz składniowa, indeksacja, tłumaczenie z lub na języki obce. Stopa błędów wynosząca zwykle kilka procent, przede wszystkim wynika z trudności modelowania obrazu rzeczywistości pozajęzykowej (poziom semantyczny). Mimo tych niedogodności automatyczne przetwarzanie tekstów jest dziedziną rozwojową, gdyż maszyna jest szybsza od człowieka i może w miarę szybko przetworzyć teksty tak duże, że człowiekowi nie starczyłoby życia na wykonanie analogicznych działań. Automatyczne tłumaczenie tekstów umożliwia np. osobom nie znającym języka chińskiego

zorientowanie się w treści dokumentów wytworzonych w tym języku, a to może służyć wstępnej selekcji dokumentów. Natomiast po tłumaczenie wysokiej jakości można i trzeba zwrócić się do prawdziwego (osobowego) tłumacza. Automatyczne indeksowanie tekstów z kolei umożliwia wyszukiwanie realizacji zadanych jednostek językowych i ich kontekstów, a nawet wyszukiwanie dokumentów zawierających ich wystąpienia.

Działania typu językowego można prowadzić na ciągach kodów znaków (liter) zawartych w pliku przechowującym (lub pozyskanych z pliku przechowującego) stronę internetową lub dokument. Natomiast obrazy zapamiętane w tych plikach do działań typu językowego nie nadają się, chyba że obraz zawiera obraz tekstu, wtedy można poddać go procedurze automatycznego rozpoznawania, a wynik umieścić w pliku źródłowym lub pliku utworzonym specjalnie do jego przechowywania. Programy rozpoznające tekst, czyli przekształcające obraz tekstu na ciąg kodów znaków (liter), nazywane są OCR-ami (Optical Character Recognition).

Pliki TXT zawierają jedynie kody znaków tekstu. Za pomocą edytora można w nich wyszukiwać zadane ciągi kodów znaków oraz je kopiować. Do wyświetlania na ekranie komputera kodów znaków tekstu używany jest wybrany, jeden dla całego dokumentu, zestaw czcionek spośród zestawów zainstalowanych na komputerze.[7]

Pliki DOC zawierają kody znaków tekstu wraz z opisem strony i opisem sposobu wyświetlania każdego znaku branego z osobna, a ponad to mogą zawierać obrazy. Za pomocą edytora można wśród kodów znaków tekstu wyszukiwać zadane ciągi kodów znaków oraz je kopiować. Każdy kod znaku tekstu może być wyświetlony za pomocą czcionki, pod względem kroju indywidualnie dla niego wybranej spośród zestawów czcionek zainstalowanych na lokalnym komputerze.

Pliki XML zawierają kody znaków tekstu wraz z opisem struktury tekstu. Obrazy znajdują się w osobnych typowych plikach graficznych jak np. BMP, JPG. Do selekcji tekstu i wyświetlania służą specjalne pliki jak np. pliki XSL, XSL-FO, XSLT. Do edycji tych plików można używać dowolnych edytorów zdolnych zapisać plik tekstowy, może to być nawet edytor tak prosty jak Notatnik. Do przeprowadzania testowych selekcji informacji zawartej w otagowanych dokumentach można używać specjalistycznego edytora XML Notepad.

Pliki DjVu (djvu lub djv) zawsze zawierają zeskanowany lub sfotografowany obraz dokumentu, który jest skompresowany w sposób dostosowany do kompresji obrazów tekstów i służy do wyświetlania dokumentu na ekranie. Pliki te opcjonalnie mogą zawierać kody znaków tekstu, najczęściej pochodzące z automatycznego rozpoznania programem typu OCR. Wtedy można kopiować kody znaków tekstu lub dokonywać wyszukiwań tekstu za pomocą wyszu-

[7] Zestawy te są instalowane za pomocą polecenia Start | Panel sterowania | Opcje regionalne i językowe | Języki | Szczegóły.

kiwarki skojarzonej z dokumentem, a same dokumenty traktować jako dwuwarstwowe, czyli składające się ze skojarzonych warstw: warstwy obrazu i warstwy kodów znaków tekstu.

Pliki PDF mogą zawierać obszary składające się z obu warstw, czyli warstwy obrazu i skojarzonej z nią warstwy kodów znaków tekstu, albo zawierać obszary składające się z tylko jednej z tych warstw. Warstwa obrazu służy do wyświetlania, a warstwa kodów znaków tekstu służy do wyszukiwania zadanego tekstu w postaci niezbyt długiego wzorca oraz do kopiowania wybranych fragmentów tekstu. Obszary nie zawierające warstwy obrazu są wyświetlane na podstawie kodów znaków i zestawów czcionek zapamiętanych w dokumencie, a jeśli danego zestawu w dokumencie nie zapamiętano, to używane są zestawy czcionek zainstalowane na lokalnym komputerze.

We wspomnianych typach plików kody znaków tekstu mogą pochodzić z ręcznego wpisania za pomocą klawiatury, z kopiowania przez schowek, lub z automatycznego rozpoznawania tekstu za pomocą programu OCR, co szczególnie dotyczy plików DjVu i PDF. Błędy ręcznego wpisywania najczęściej wynikają z nieuwagi osoby używającej klawiatury. Błędy kopiowania przez schowek mogą pochodzić z niezgodności stron kodowych lub zmiany formatu tekstu. Natomiast błędy automatycznego rozpoznawania mogą pochodzić ze złej jakości druku dokumentu źródłowego, źle dobranych warunków skanowania lub braku rozpoznawanego wyrazu w podręcznym słowniku programu rozpoznającego, co zwykle dotyczy wyrazów rzadko używanych lub wyrazów nowych, które wystąpiły w tekście rozpoznawanym. W tym ostatnim przypadku program rozpoznający zwykle zamienia wątpliwy wyraz innym wyrazem znajdującym się w słowniku podręcznym programu, z jednoczesnym zaznaczeniem wątpliwego wyrazu tak, że przed zapamiętaniem wyniku rozpoznawania można ręcznie poprawić wynik rozpoznawania.

Dosyć częstym błędem jest zapamiętanie obrazu dokumentu pierwotnego z użyciem zbyt małej zdolności rozdzielczej urządzenia skanującego lub fotografującego w celu oszczędności miejsca w pamięci stałej komputera (dyski twarde), zdarza się to nawet w czołowych bibliotekach polskich. Dokonywane następnie automatyczne rozpoznawanie pisma jest obciążone zbyt dużą liczbą błędów, co w konsekwencji utrudnia lub uniemożliwia działania językoznawcze badaczom stosującym metody informatyczne, dla których materiał wejściowy musi być bezbłędny, gdyż komputery działają szybko, ale niczego się nie domyślą ponad to, co w programie zawarł programista. Dla szerokiego zakresu różnych typów dokumentów bezpieczną zdolnością rozdzielczą skanowania jest 600 dpi, z wyjątkiem szczególnie cennych dokumentów, dla których trzeba stosować skanowanie ze zdolnością rozdzielczą od 3000 do 6400 dpi.

Jeszcze poważniejszym błędem jest zeskanowanie ze zbyt małą zdolnością rozdzielczą dokumentów, które należałoby zachować, a następnie pozbycie się ich papierowych oryginałów, gdyż wtedy powstaje sytuacja trudna lub niemożliwa do naprawienia. Natomiast, gdy papierowe oryginały zostały zachowane, to

można mieć nadzieję, że w przyszłości znajdzie się ekipa osób kompetentnych i fachowych, która zdoła naprawić błędy poprzedników. Przy okazji warto wspomnieć, że dokumenty papierowe mają większą wartość dowodową niż dokumenty elektroniczne oraz że właściwie przechowywane dokumenty papierowe są bardziej trwałe, a ich przechowywanie jest mniej kosztowne niż przechowywanie dokumentów elektronicznych. Z kolei wielką zaletą dokumentów elektronicznych jest możliwość szybkiego i taniego przesyłania ich na duże odległości oraz możliwość przetwarzania metodami językoznawczo-informatycznymi, a w tym szybkiego wyszukiwania pożądanych treści.

Działanie popularnej wyszukiwarki stron internetowych Google jest w zarysie wyjaśnione na stronie internetowej *Indeksowanie – Wszystko o wyszukiwaniu*, gdzie stwierdzono „Nasz indeks ma sporo ponad sto milionów gigabajtów, a na jego tworzenie poświęciliśmy ponad milion godzin obliczeń". Wyszukiwarka Google indeksuje tzw. wyrazy tekstowe „od spacji do spacji" uwzględniając jako miejsca delimitacji znaki przestankowe oraz niektóre znaki spoza alfabetu. Mimo tak prostego i mechanicznego indeksowania, do pewnego stopnia obarczonego błędami, wyszukiwarka jest ważnym narzędziem językowym z uwagi na wielką liczbę zaindeksowanych stron.

Andrzej Wawrzyńczyk porównuje działanie wyszukiwarek Google, Onet.pl Szukaj, NetSprint.pl oraz Szukacz (Wawrzyńczyk 2006). Wyróżnia on następujące rodzaje pytań: *Pytanie o słowo*, *Pytanie o fragment słowa*, *Pytanie o wyrażenie* oraz *Ograniczenie liczby wyników*. Przy tym zwraca uwagę na próby uwzględniania polskiej fleksji oraz stosowania znaków wieloznacznych. Ponadto proponuje nazwać polskojęzyczne strony dostępne za pomocą tych wyszukiwarek Polskim Korpusem Internetowym, w skrócie PKI. Jest to ciekawa propozycja, gdyż skłania do refleksji nad pytaniem o właściwości korpusów oraz nad pytaniem „w jakim stopniu PKI jest korpusem?". Ponadto A. Wawrzyńczyk omawiając istniejące korpusy instytucjonalne przy niektórych nadmienia, że korpus „jest w dalszym ciągu rozwijany" (Wawrzyńczyk 2006: 16), „Wydawnictwo ... tworzy i na bieżąco wykorzystuje własny korpus języka polskiego" (Wawrzyńczyk 2006: 19), czyli zmienia się zawartość korpusu.

Językoznawstwo powinno być w swej istocie uprawiane jako dyscyplina doświadczalna w taki sposób, by różni badacze mogli wzajemnie porównywać wyniki doświadczeń. W szczególności drugi badacz, weryfikujący wyniki badacza pierwszego, powinien w tych samych warunkach, uzyskać ten sam wynik co badacz pierwszy dla tej samej hipotezy badawczej, metody weryfikacji hipotezy i tego samego materiału wejściowego, czyli korpusu. Zatem stałość korpusu jest ważnym czynnikiem rzetelnie przeprowadzonego doświadczenia. Korpus powinien być dostępny dla badacza w postaci niezaszyfrowanej, aby mógł on obejrzeć każdy wybrany dowolnie duży fragment korpusu. Dwie realizacje korpusu są na pewno identyczne, gdy są identyczne bajt w bajt. Jeśli korpus instytucjonalny lub prywatny jest w budowie, to może być używany do porównywalnej, i przez to wiarygodnej, weryfikacji hipotez przez wielu badaczy

pod warunkiem, że każda opublikowana realizacja korpusu oprócz nazwy będzie oznaczona wersją realizacji, aby w razie uzyskania różnych wyników było jasne, że różnica wyników nie powstała na skutek różnic między wersjami korpusów, używanymi przez różnych badaczy. Zupełnie inną sprawą jest to, że pozytywnie zweryfikowana hipoteza badawcza (prawo językowe) jest w pewnym sensie tym mocniejsza, im dla szerszego spektrum korpusów daje ten sam wynik.

W odpowiedzi na zapytanie wyszukiwarka Google standardowo podaje adresy pierwszej dziesiątki stron o najwyższym priorytecie. Można ręcznie wywoływać następne dziesiątki stron, aż do wyczerpania zestawu adresów stron dostarczonych w odpowiedzi lub aż do dojścia do limitu, który zależnie od komputera wynosi około tysiąca stron, nawet jeśli wyszukiwarka Google podała, że znalazła kilkaset milionów wyników. Większość użytkowników ogląda co najwyżej pierwszą dziesiątkę zaanonsowanych stron, zatem specjaliści zwani pozycjonerami starają się uzyskać wyższy priorytet dla stron znajdujących się pod ich pieczą. W jaki sposób nowa strona ma być pobrana, zaindeksowana i zapamiętana przez Google, webmaster określa w pliku robots.txt. Żądanie nieindeksowania strony powoduje, że nie jest ona uwzględniana w oficjalnym indeksie Google. Z kolei strony rzadziej odwiedzane przez internautów oraz mające mniej powiązań z popularnymi stronami, mogą zostać automatycznie usunięte z oficjalnego indeksu Google, tak samo jak strony usunięte z serwera przez właściciela strony. Strona nieuwzględniona w oficjalnym indeksie Google nie może wystąpić w odpowiedzi na zapytanie i w tym sensie nie istnieje. Aby do takiej, pozornie nie istniejącej strony dostać się, trzeba znać jej adres lub adres jej witryny i sposób dojścia do niej na witrynie.

Wyszukiwarka Google i inne wyszukiwarki internetowe są ważnymi narzędziami dla językoznawcy, można ich używać do wyszukiwania stron zawierających zadane z góry wyrazy lub wyrażenia, a z uwagi na liczbę potencjalnie dostępnych stron mogą symulować korpusy językowe. Strony uwzględnione w ich indeksach prawdziwymi korpusami jednak nie są i nie będą, ze względu na swoją zmienność i inne wyżej wskazane ograniczenia.

Wśród językoznawców zainteresowanych badaniem tekstów dostarczanych łączami internetowymi wyróżniają się dwa typy. Pierwszy, umownie nazywam go typem A, selekcjonuje, czyta i analizuje dostarczane teksty. Nawigowanie po witrynach o skomplikowanej strukturze odbiera jako męczące. Najchętniej widziałby tekst jako długą, przewijaną taśmę. W momencie znalezienia interesującej jednostki, chciałby łatwo uzyskać dane bibliograficzne miejsca jej wystąpienia oraz móc łatwo skopiować (udokumentować) znalezisko. Typ drugi, nazywam go typem B, selekcjonuje i gromadzi dostarczane teksty. W zasadzie ich nie czyta, lecz układa testy automatycznego sprawdzania. W czasie układania testu wyrywkowo czyta badany tekst, sprawdza również konteksty wyników.

Dla obu typów badaczy jednym z podstawowych pytań jest pytanie, czy badane teksty należy sprowadzić na lokalny komputer. W razie pozytywnej

odpowiedzi warto pamiętać, że strony internetowe zawierające ilustracje (obrazy) oraz zagnieżdżone dokumenty zawierające ilustracje (obrazy) zajmują wielokrotnie więcej miejsca od stron internetowych i dokumentów nie zawierających ilustracji.

Od pewnego czasu działają grupy entuzjastów tworzących pliki z rekordami katalogującymi strony internetowe, które można przeszukiwać za pomocą języka SPARQL[8] będącego odmianą języka SQL. Zapewne te doświadczenia można by wykorzystać przy budowie narzędzi dla językoznawców. Język SPARQL ma możliwość testowania dopasowania wzorca z zapytania nie tylko do zawartości całych pól, ale również do ich części. Rekord opisujący stronę internetową, jako podstawowa jednostka stronografii internetowej, powinien zawierać adres strony oraz podstawowe dane zaczerpnięte z kodu źródłowego strony, z części nazywanej nagłówkiem, gdzie zwykle znajdują się informacje takie jak *Tytuł strony*, *Wyrazy kluczowe*, *Język strony*, *Autor strony*, *Utrata ważności* i inne. Jeśli na stronie są zagnieżdżone dokumenty (DOC, PDF, DjVu itp.), to w rekordzie opisującym stronę powinny znaleźć się podstawowe dane dotyczące tych dokumentów. Jeśli dokumenty są kopiami druków zwartych lub czasopism, to dane można zaczerpnąć z klasycznych opisów bibliograficznych. Kłopot w tym, że webmasterzy i pozycjonerzy często w nagłówkach stron internetowych umieszczają informacje nieadekwatne do treści strony, a nawet informacje fałszywe, gdyż starają się podnieść rangę strony w indeksie Google. Zatem przed opisaniem trzeba obejrzeć opisywaną stronę internetową; szczególnie, gdy osoba opisująca nie jest pewna wiarygodności twórcy oraz właściciela strony internetowej.

Oprócz opracowania pliku z rekordami opisującymi strony internetowe wraz z dokumentami ewentualnie na nich znajdującymi się, potrzebne by było opracowanie schematów zapytań selekcjonujących dokumenty żądanego rodzaju, interfejs użytkownika i mechanizm dokumentujący znalezisko lub nawet dodatkowo całe badane strony internetowe.

Narzędzie (system pozyskiwania i przetwarzania tekstów) powinno móc pracować co najmniej w trzech podstawowych trybach:

1. trybie wyświetlania obrazu strony internetowej lub obrazu zagnieżdżonego dokumentu;
2. trybie wyświetlania kodów znaków strony internetowej jak i kodów znaków zagnieżdżonego dokumentu (o ile taki jest);
3. trybie automatycznego testowania wzorców (automatycznego wyszukiwania jednostek językowych) w kodach znaków tekstu.

[8] Opisy języka SPARQL można znaleźć na stronach:
http://www.w3.org/2009/Talks/0615-qbe/Overview.html
http://www.iro.umontreal.ca/~lapalme/ift6281/sparql-1_1-cheat-sheet.pdf
http://www.slideshare.net/LeeFeigenbaum/sparql-cheat-sheet
a specyfikację języka na stronie:
www.w3.org/TR/sparql11-query

Jeśli na stronie internetowej znajduje się dokument taki, jak np. książka, to można się spodziewać, że oglądany tekst jest tekstem spójnym. Natomiast gdy oglądane są po prostu strony internetowe nie zawierające zagnieżdżonych dokumentów, a witryna ma skomplikowaną strukturę typu sieci z cyklami, to można się spodziewać, że po linearyzacji tej struktury, tekst miejscami, tam gdzie zostały rozcięte łącza, może stracić spójność.

Językoznawcę typu B interesuje możliwość badania dużych ilości tekstów, które są do pewnego stopnia jednorodne, gdyż narzędzia służące do analizy dobiera on do formatu tekstów. Jeśli zbiór badanych stron internetowych zawiera zagnieżdżone dokumenty tego samego formatu (np. DOC, PDF lub DjVu), wtedy warto wyłuskać te dokumenty ze stron internetowych, a nawet jeśli się da, to wyłuskać warstwy tekstowe z dokumentów i złożyć je w osobnych plikach. Każda z tych operacji ułatwia i przyspiesza dalsze badania. Dla potrzeb badacza typu B warto zintegrować narzędzie (system) służące do analizy stron internetowych z parserem wyrażeń regularnych oraz z kompilatorami języków takich, jak Perl czy Pascal. Pożyteczne będzie również narzędzie wyłuskujące kody tekstu widocznego na stronie internetowej. Wyrażenia regularne mogą służyć do konstruowania filtrów opisanych np. w pracy *Filtry Wierzchonia jako narzędzie badawcze filologa* (Małek 2006). Natomiast parser wyrażeń regularnych może służyć do wyszukiwania realizacji konstrukcji składniowych takich jak opisane w pracy (Wierzchoń 2008). Wyrażenia regularne są narzędziem względnie szybko pracującym, niemniej im dłuższe wyrażenie tym staje się mniej czytelne, np. wyrażenia podane na stronach 295 i 310 w książce Jeffreya Friedla (Friedl 2001).

Jeśli korpus jest udostępniany badaczom w trybie on-line w taki sposób, że nie mogą oni przeglądać całości np. za pomocą przewijania tekstu lub wydawania instrukcji *Pokaż następną jednostkę*, to taki korpus nie nadaje się do poważnych badań.

W niniejszej publikacji[9] starałem się zarysować problem (w stylu „instruktażowym"). Pytanie zawarte w jej tytule pozostaje aktualne. O rozwiązanie można by poprosić informatyków, specjalistów od Internetu i korpusów językowych. W wielu instytutach o profilu językoznawczym lub informatycznym istnieją już komórki zajmujące się implementowaniem metod informatycznych. Zdaje się, że jest szansa na zebranie zespołu entuzjastów.

[9] W pracy nad nią wykorzystałem notatki do referatu (pt. „Filolog pyta o bibliografię zawartości Internetu"), który wygłosiłem w dniu 4 X 2013 r. na międzynarodowej konferencji naukowej w Instytucie Slawistyki PAN (nazwa konferencji: „Nowoczesne systemy slawistycznej informacji bibliograficznej – dziś i jutro"); współorganizatorami konferencji były następujące placówki naukowe: Instytut Slawistyki Zachodniej i Południowej UW, Fundacja Slawistyczna, Towarzystwo Naukowe Warszawskie oraz Komisja Bibliografii Lingwistycznej przy Międzynarodowym Komitecie Slawistów.

Literatura

Bomba R., 2013. *Narzędzia cyfrowe jako wyznacznik nowego paradygmatu badań humanistycznych*. – [In:] Radomski A., Bomba R. (red.), *op. cit.*

Friedl J. E. F., 2001, *Wyrażenia regularne*. Przeł. A. Podstawczyński. – Gliwice.

Małek E., 2006, *Filtry Wierzchonia jako narzędzie badawcze filologa*. – Łódź.

Radomski A., Bomba R. (red.) 2013, *Zwrot cyfrowy w humanistyce. Internet / nowe media / kultura 2.0*. – Lublin.

Wawrzyńczyk A., 2006, *Korpusy językowe. Tekstowe zasoby Internetu jako korpus. Wprowadzenie*. – Warszawa.

Wawrzyńczyk J., 1998, *Informacyjne wspomaganie polonistyki językoznawczej (uwagi użytkownika bibliografii)*. – [In:] "Praktyka i Teoria Informacji Naukowej i Technicznej" (Warszawa), t. VI, nr 3-4, s. 5-10.

Wawrzyńczyk J., 2000-2012, *Słownik bibliograficzny języka polskiego. Wersja przedelektroniczna*, t. 1-10. – Warszawa.

Wierzchoń P., 2008, *Kotuś : "Verba polona abscondita..." (w fotodokumentacji) : szkic lingwochronologizacyjny : centuria pierwsza*. – Poznań.