Eliza Małek

FILTRY WIERZCHONIA JAKO NARZĘDZIE BADAWCZE FILOLOGA

Łódź 2006

Recenzent:
Piotr K. Witas

Wydawca:
Instytut Rusycystyki Uniwersytetu Łódzkiego

ISBN 83-60416-99-0

Wyd. I. Obj. 0,18 a.w.

WSTĘP

Kilka lat temu w serii „Semiosis Lexicographica", wydawanej na Uniwersytecie Warszawskim, ukazała się (jako jej tom 15.) niewielka objętościowo praca Piotra Wierzchonia pod zaskakującym tytułem: *Z cudzysłowów do poczekalni leksykograficznej* [Wierzchoń 2003]; rok temu ten młody poznański lingwista i informatyk w jednej osobie, kierownik Zakładu Koreanistyki Uniwersytetu im. Adama Mickiewicza, opublikował jej część drugą, obszerniejszą [Wierzchoń 2005; por. tejże serii tom 26.].

Można tu mówić o wydarzeniu naukowym, jakim jest przedstawione w tych dwu publikacjach narzędzie badawcze, nazwane przeze mnie filtrami Wierzchonia.

Dla filologów (tradycyjnych i nietradycyjnych, „ponowoczesnych", badaczy literatur, języków), ale i niefilologów (badaczy kultur, publicystów, dziennikarzy, pisarzy itp.), wszystkich, którzy mają do czynienia ze słowem, słowem-przedmiotem, słowem-narzędziem, są tego słowa aktywnymi użytkownikami, a nierzadko i twórcami, dla żyjących w świecie słów, zdań, w świecie tekstów sprawą zasadniczą jest przecież sprawne i szybkie poruszanie się po tym świecie. Jesteśmy świadkami narodzin cyfrowej „galaktyki Gutenberga", coraz więcej tekstów powstaje nie na papierze, lecz w wersji (wyłącznie) elektronicznej, bądź równolegle, na papierze i w postaci cyfrowej. Nabiera tempa digitalizacja zasobów bibliotecznych, ratująca je przed zniszczeniem, rozpadaniem się kwaśnego papieru. Być może już dzisiaj tekstów na papierze jest mniej niż tekstów na nośnikach elektronicznych. Po oceanach tekstów trzeba się nauczyć żeglować. Tych podróży nie boją się młodzi. Starsze pokolenia reagują różnie (znam profesorów nadal nie używających komputera do pisania swoich prac, nie umiejących obsługiwać poczty elektronicznej).

Część filologów wciąż jeszcze wypisuje z badanych tekstów potrzebne „byty graficzne" na fiszki, układa kartoteki w pudełkach czy szufladkach, nieświadoma, że istnieje coś takiego, jak wspomaganie komputerowe, informatyczne badań filologicznych (szerzej: humanistycznych). Te tradycyjne zachowania już są nieodwracalnie anachroniczne, dla młodzieży filologicznej (humanistycznej) zupełnie nie do przyjęcia. Filolog, badacz wrażliwy na jakość swojej pracy[1] nie może nie docenić zalet i wielkich przewag cyfrowej „rewolucji". *To idzie e-młodość!*

[1] Kierujący się określoną metodologią badawczą; w tym momencie niech mi zarazem wolno będzie przypomnieć ostrzegawczo dawną, złośliwą, ale celną i aktualną uwagę Franciszka Salezego Dmochowskiego (z 1858 r.): „zwyczajem wszystkich filologów więcej przywiązywał się do słów, niżeli do gruntu rzeczy" [za: Wawrzyńczyk 2004, s.v. **filolog**].

Przyrastająca z błyskawiczną szybkością masa e-tekstów, zwłaszcza w Internecie (w tym w Runecie[2]), jest nieporównywalnie łatwiej dostępna czytelniczo niż teksty na papierze (a e-oporny stary filolog z dziecinną łopatką do piaskownicy w ręce – chce rozkopać Mount Everest papierowy...). Choć z drugiej strony: gwarancja jakości merytorycznej i formalnej, edytorskiej tekstów papierowych jest większa niż w wypadku e-tekstów; w Internecie każdy może wstawić swoją pracę, nie ma tu redaktorów, kolegiów redakcyjnych, recenzentów (z wyjątkiem, częściowym, czasopism elektronicznych).

Znaczenie Internetu, do którego trafia coraz więcej literatury naukowej, jak i samych źródeł przydatnych badaczom, nieustannie rośnie. Brak nawyku zaglądania do Internetu, korzystania z jego zasobów − oczywiście jak wypada każdemu solidnemu badaczowi, korzystania z maksymalną dozą krytycyzmu − jest błędem.

Pożytek, jaki przynosi dzisiaj Internet, można zilustrować paroma przykładami z językoznawstwa[3]. Np. dzisiaj ustalenia i wnioski badawcze z zakresu lingwistyki barw zawarte w publikacji sprzed 12 lat [Ampel-Rudolf 1994] muszą w znacznej części, dzięki materiałowi tekstowemu dostępnemu elektronicznie, zostać znacznie rozbudowane. W ogóle niektóre tematy muszą być podjęte na nowo, niektóre prace napisane niejako od nowa. Szczególnie przejawi się ten „e-przymus" rewizji na obszarze słowotwórstwa, frazeologii, leksykografii, w badaniach historii słownictwa polskiego czy rosyjskiego (by wymienić języki mi najbliższe)[4].

Niewątpliwie także badacz literatury czy historyk kultury (etnokultury) współcześnie ma do dyspozycji znacznie więcej danych niż jeszcze kilka czy kilkanaście lat temu. Ktoś, kto kiedyś pisał o takich przedmiotach kultury materialnej, o ich symbolice w tekstach artystycznych, jak *баня*, *велосипед*, *самовар*, *трамвай*, *телефон*,[5], dzisiaj musiałby znacznie rozszerzyć swoją wiedzę na te tematy, sięgnąć do zupełnie nowych źródeł, w tym e-źródeł.

E-źródła zaś wymagają filtrowania.

---

[2] Zdaje się on zawierać więcej tekstów interesujących filologa, zwłaszcza utworów literatury pięknej, niż polska sekcja Światowej Sieci.

[3] Mówi się już o e-lingwistyce.

[4] Np. w rusycystyce językoznawczej zupełnie brak prac weryfikujących (falsyfikujących) datacje, chronologię słownictwa opisywanego w ramach wielkiego akademickiego cyklu publikacji pod „przechodnim" tytułem *Новое в русской лексике* i *Новые слова и значения*. Jan Wawrzyńczyk poinformował mnie, że niektóre jednostki, określane tam jako nowe na podstawie wystąpień w tekstach prasowych z lat 60-tych (i późniejszych) ubiegłego wieku, znajdują się w cytatach zawartych w słownikach języka rosyjskiego z tychże lat 60-tych i wcześniejszych; chodzi tu o wyrazy *de facto* ukryte w owych cytatach ilustracyjnych, bo nie umieszczone w siatce haseł tych słowników.

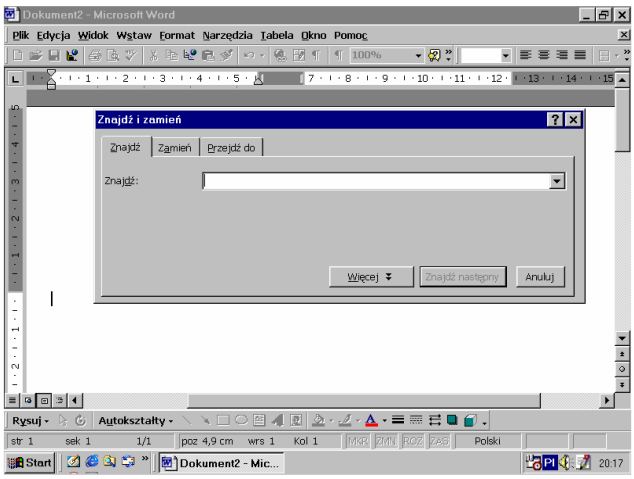
[5] Por. te hasła w [Wawrzyńczyk, Małek 2004].

O FILTROWANIU TEKSTÓW

Filtry to − jak wynika z lektury[6] odkrywczych prac Piotra Wierzchonia, przywołanych w niniejszej broszurze − nic innego jak wymyślone *ad hoc* określenie prostego zapisu formuł napisanych w języku wyrażeń regularnych (RE). Używa się tego określenia z braku jakiegoś innego, który by się wydał odpowiedniejszy, stosowniejszy (bardziej fonoestetyczny?), jest zatem właściwie obojętne, jakie będą jego (tego terminu) dalsze losy. Filtry mają służyć przede wszystkim rozwiązaniom praktycznym: mają pomagać znaleźć w tekście określone fragmenty napisów, zapisów, wyrażeń zatem graficznych, grafemowych. Przeto jeżeli chcemy odszukać w tekście słowo **`domek`**, to wpisujemy w dowolny program[7], który obsługuje składnię wyrażeń regularnych, napis **`domek`**.

[6] Nie dla każdego literaturoznawcy łatwej.

[7] To może być pierwsza komunikacyjna niejasność. Programów obsługujących wyrażenia regularne są setki. Każdy bardziej zaawansowany program programistyczny (służący do pisania programów lub np. stron www) obsługuje ten mechanizm. Co więcej, nawet w pewnym zakresie obsługuje ten mechanizm M$ Word. Wystarczy w polu Znajdź zaznaczyć: *użyj symboli wieloznacznych* i już można wpisywać formuły: [a-z], [0-9] itd. Oznacza to kolejno: znajdź dowolną literę od a do z, znajdź dowolną cyfrę od 0 do 9. Natomiast zapis [0-9]+ oznacza: znajdź dowolny ciąg cyfr, czyli *de facto* jakąkolwiek liczbę.



Naturalnie, jeżeli chcemy wyszukać wyraz **domek**, nie musimy mieć programu obsługującego RE, bo siła szukania w RE wynika z szukania inwariantów graficznych. Zatem szukamy niezmiennej formy graficznej: `dom, domem, domy,` (w tym ciągów) `małe domy, małymi domami` itd. Chodzi wobec tego o odpowiednie sformułowanie takiego wersu poszukiwania, by obejmował on swą postacią maksymalną liczbę przypadków (tj. postaci graficznych, np. wynikłych z morfologii, ze zjawisk fleksyjnych), które nas w danej chwili interesują. Dlatego przykładem bardziej zaawansowanego filtra jest postać: domk[i|iem|u]. Zapis ten oznacza: wyszukaj w tekście: **domki** lub **domkiem**, lub **domku**.

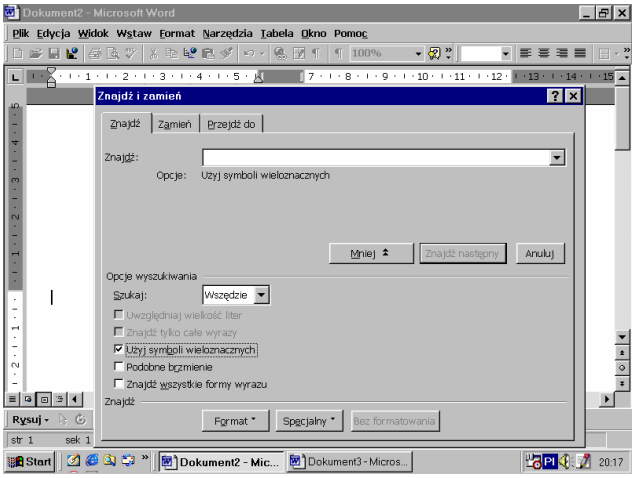
Na przykład, jeżeli interesuje nas wyraz z łącznikami, to piszemy jako inwariant:

[a-z]+-[a-z]+-[a-z]+

co oznacza: znajdź wyraz złożony z jakichkolwiek liter (a-z) i dowolnie długi (to gwarantuje plusik), potem łącznik i potem dwa razy taką samą sytuację, biorąc za inwariant jakikolwiek wyraz, po którym jest łącznik. Rzecz w tym, żeby w tekście znajdować to, czego potrzebujemy. Aby to zrobić, trzeba określić maksymalny inwariant graficzny (czyli to, co się nie zmieni, np. że nie ulegną zmianie łączniki w wyrazie trójłącznikowym).

Konstruowanie konkretnych filtrów jest uwarunkowane konkretnymi życzeniami lingwisty, dlatego można te filtry pisać bez końca, np.: znajdź wszystkie wyrazy rozpoczynające się od anty-:

anty*
lub
anty[a-z]+
lub



anty.*

Skąd te różnice? Otóż jak to w świecie informatyków się zdarza, nie ma jednego standardu kodowania wyrażeń regularnych. Stąd więc różne programy (por. przypis 1.) kodują różnie te sytuacje tekstowe. Wynika to po prostu już z samego mechanizmu danego programu.

Kluczowa idea filtrowania została zastosowana już w pracy o cudzysłowach [Wierzchoń 2003]. Tam poszukiwane były jednostki:

"[a-z]+"

a więc jednostki, przed którymi (oraz po których) pojawiał się cudzysłów.

Oczywiście o powyższych przykładach trudno mówić, że są wyrafinowane pod względem lingwistycznym. Istota pomysłu filtrów wynikała pierwotnie z chęci wyszukania w miarę stałych połączeń wyrazowych (wyszukiwanie kolokacji). Ponieważ język polski jest językiem fleksyjnym, należało zaproponować jakiś bardziej prymitywny od światowego (tj. dla angielskiego) mechanizm (kwantytatywne liczenie wszystkich par w tekstach).

W [Wierzchoń 2002] filtry ujęte zostały w następujący sposób:

"Na przykład chcemy odnaleźć połączenie wyrazowe występujące po wyrazie: “przezwano”, a jednocześnie interesuje nas potencjalne wystąpienie takich połączeń po ciągach: *przezwano go*, *przezwano ich*, *przezwano je*, *przezwano ją*, *przezwano to*. Formułujemy zatem jedno wyrażenie:

`przezwano (go|ich|je|ją|to|) [a-˙]+ [a-˙]+"`

Formuła ta zatem pozwala użytkownikowi („filtratorowi”) odnaleźć wszystkie ciągi dwuwyrazowe, które poprzedzono informacją *przezwano* oraz *go*, *ich*, *je*, *przezwano ją*, *to*.

W artykule [Wierzchoń 2002] autor skoncentrował się na następujących filtrach zawierających ciągi: *nazwano*, *określa się*, *określa się mianem*, *nazywa się*, *tzw*. Tamże czytelnik znajdzie propozycję konstrukcji poszczególnych filtrów oraz omówienie problemów i kłopotów powstających podczas pracy z konkretnym filtrem. Przedstawione zostały ilustracyjnie filtry:

1. nazwan[a-˙]+
2. nazwano (go|ich|je|jego|jej|ją|to)
3. nazwan[a-˙]+ by
4. nazwano by (go|ich|je|ją|to)
5. nazwan[a-˙]+ przez [a-˙]+

6. nazwan[a-˙]+ został+
7. nazywa się
8. nazywa się (go|ich|je|ją|on|ona|ono|to)
9. nazywa się (także|też)
10. nazywa się (czasem|czasami)
11. nazywa się (potocznie|inaczej|po prostu)
12. mianem
13. mianem tym określ[a-˙]+
14. określa się
15. określa się (go|ich|je|to)
(czasem|czasami|często|także|zwykle|niekiedy|nawet) jako
16. tak zwan[a-˙]+
17. tzw\.

Po takich ciągach możliwe jest wprowadzenie dotyczące
dwu- `[a-˙]+ [a-˙]+`
lub więcejwyrazowych `[a-˙]+ [a-˙]+ [a-˙]+`
połączeń (lub jednego wyrazu `[a-˙]+`).

Autor pracuje nad udoskonaleniem swoich propozycji, zmierzającym do pełniejszej i efektywniejszej automatyzacji ekscerpcji połączeń wyrazowych.

BIBLIOGRAFIA

Ampel-Rudolf, Mirosława (1994). *Kolory. Z badań leksykalnych i składniowo-semantycznych języka polskiego*, Rzeszów: WSP.

Bańczerowski, Jerzy (ed.) (1991). *The application of microcomputers in the humanities*, Poznań: UAM.

Dudzińska, Aleksandra (2005). *Język rosyjski w Internecie. Zarys problematyki*, Warszawa: Semiosis Lexicographica.

Wawrzyńczyk, Andrzej (2006). *Korpusy językowe. Tekstowe zasoby Internetu jako korpus. Wprowadzenie*, Warszawa: Takt.

Wawrzyńczyk, Jan (2004). *Słownik bibliograficzny języka polskiego. Wersja przedelektroniczna. T. 2: D-G*, Warszawa: Semiosis Lexicographica.

Wawrzyńczyk, Jan, Małek, Eliza (2004). *Z materiałów do Słownika bibliograficznego języka rosyjskiego. Terminologia lingwistyczna. Wybrane terminy wiedzy o kulturze i literatureze. Neologizmy, hapaks legomena*, Warszawa: Semiosis Lexicographica.

Wawrzyńczyk, Jan (red.) (2004). *Korpusy języka rosyjskiego w Polsce i na świecie*, Warszawa: Semiosis Lexicographica.

Wierzchoń, Piotr (2002). *Automatyzacja ekscerpcji definiowanych połączeń wyrazowych. Filtry wyrażeń regularnych*, [w:] Krzemińska, W., Nowak, P. (red.), *Przestrzenie informacji*, Poznań: Sorus, s. 119-184.

Wierzchoń, Piotr (2004). *Gramatyka diakrytologiczna. Studium ortograficzno-kwantytatywne*, Poznań: Wydawnictwo UAM.

Wierzchoń, Piotr (2003). *Z cudzysłowów do poczekalni leksykograficznej*, Warszawa: Semiosis Lexicographica.

Wierzchoń, Piotr (2005). *Z cudzysłowów do poczekalni leksykograficznej. II*, Warszawa: Takt.

Notatki

SPIS TREŚCI