Krzysztof Szafran, Mirosława Miłkowska, Ewa Zakrzewska

Z doświadczeń pracy nad konkordancjami do Vade-mecum Cypriana Norwida [1]

0. Wstęp

W październiku 1985 roku został sporządzony próbny tekst konkordancji do „Vade-mecum" Cypriana Norwida (w dalszym ciągu niniejszego artykułu pisać będziemy Konkordancje — mając na myśli konkordancje do „Vade-mecum" C. Norwida). Tym samym zakończony został pewien etap prac zespołu zajmującego się komputerowym opracowaniem tekstów Norwida. Prace te zostały zainicjowane przez istniejącą przy Uniwersytecie Warszawskim (dalej UW) Pracownię Słownika Języka Cypriana Norwida (dalej PSJCN), kierowaną przez prof. dr hab. Jadwigę Puzyninę. Dostarczyły one wielu doświadczeń wszystkim biorącym w tych pracach udział, zarówno norwidologom, jak i informatykom.

Tekst niniejszy jest opisem dość złożonego procesu, który doprowadził do powstania Konkordancji. Nie jest on jednak prezentacją poglądów całego zespołu zajmującego się ich przygotowaniem, ani tym bardziej prezentacją stanowiska PSJCN. Co więcej, niektóre przedstawione tu opinie wywołują nadal żywe dyskusje wśród autorów tekstu.

Mamy nadzieję, że poprzez opisanie naszych doświadczeń uda się nam pokazać, jak istotne w momencie podejmowania decyzji o rozpoczęciu pracy jest uświadomienie sobie problemów, które trzeba będzie rozwiązać, i to rozwiązać do końca, aż do najdrobniejszych szczegółów. Pozwala to uniknąć rozmaitych, często przykrych niespodzianek. Niestety, często się zdarza, że prace takie są prowadzone przez zmieniającą swój skład w trakcie badań grupę ludzi dość luźno związanych z zagadnieniem i wykonujących tylko zadania cząstkowe oraz ludzi współpracujących z zespołem na zasadzie krótkoterminowych prac zleconych. W takich warunkach sprawne i efektywne działanie nie jest sprawą łatwą, przede wszystkim pod względem organizacyjnym. Dlatego też bezpieczniej jest, jeśli wszelkie ustalenia podejmowane są w formie pisemnej i akceptowane przez wszystkich zainteresowanych. Przedsięwzięcie tych rozmiarów wymaga bowiem tak wielu bardzo szczegółowych ustaleń (które w dodatku zmieniają się często we wstępnej fazie pracy), że już wkrótce nikt z zainteresowanych nie pamięta dokładnie, co i jak zostało ustalone. Ponadto pomiędzy ostatecznym ustaleniem wszystkich szczegółów a uzyskaniem wyniku końcowego mija zwykle dość dużo czasu, co powoduje, że przy ocenie tego wyniku trudno jest z całą pewnością powiedzieć, jakie decyzje zostały ostatecznie podjęte. Najlepszym przykładem jest nasz zespół. Kiedy przystępowaliśmy do pisania niniejszego tekstu, stwierdziliśmy, że bardzo trudno przychodzi nam odtworzenie z pamięci niektórych istotnych szczegółów i faktów.

1. Sformułowanie zadania

W trakcie prac nad koncepcją Słownika, prowadzonych przez PSJCN, zaczęto w pewnym momencie brać pod uwagę możliwość użycia komputera. Początkowo nie było wcale zgody, czy i do czego mógłby się on przydać. Co prawda, dla informatyka pomysł wykonywania ręcznie prac typu słowników, indeksów czy konkordancji jest absurdalny, ale historia lingwistyki zna tego typu prace wykonane z powodzeniem (por. np. Sambor 1972). Ponadto wcześniejsze doświadczenia z użyciem komputerów do innych polskich prac lingwistycznych (por. Kurcz et al. 1976, Woronczak 1983) nie wzbudzały zaufania do tego narzędzia u członków PSJCN. Faktem jest, że w momencie rozważania możliwości użycia komputera do prace nad tradycyjną kartoteką z fiszkami były już bardzo zaawansowane, a ewentualne prace komputerowe miały być niezależne od prac prowadzonych metodami tradycyjnymi. Ostatecznie w wyniku wielomiesięcznych dyskusji, analiz i konsultacji PSJCN podjęła decyzję wykonania konkordancji za pomocą komputera.

Konkordancje przewidziane były jako pomoc dla uczonych — norwidologów, ale przede wszystkim jako wstępny etap prac nad Słownikiem, etap, którego głównym celem było wykazanie przydatności komputera do tego typu prac oraz sprawdzenie możliwości dostępnego oprogramowania.

Większość problemów, o których będzie mowa w dalszej części niniejszego tekstu, powinna była być, przynajmniej częściowo, rozwiązana, a rozwiązania — zaakceptowane przez PSJCN przed podjęciem decyzji o rozpoczęciu prac na

[1] Tekst niniejszy ukazał się drukiem w pracy: *Studia z polskiej leksykografii współczesnej*, tom II pod redakcją Zygmunta Saloniego, Dział Wydawnictw Filii UW w Białymstoku, 1987, ss.309-334.

komputerze. Przyjęcie takiej kolejności mogło jednak spowodować wielomiesięczne opóźnienie prac (w przypadku podjęcia decyzji pozytywnej).

W skład zespołu, który miał bezpośrednio wykonać konkordancje wchodziła część osób zatrudnionych w PSJCN oraz informatycy z Instytutu Informatyki UW (dalej IInf UW). Niestety podział kompetencji (i odpowiedzialności) w zespole nie został dostatecznie wyraźnie określony.

2. Ustalenia generalne

Zadanie, które przed nami stało, w sposób naturalny dzieli się na trzy etapy.

I. Przygotowanie tekstu na nośniku komputerowym w postaci nadającej się do dalszego przetwarzania (etap ten dzieli się z kolei na kilka istotnych podetapów, o których będzie mowa dalej).

II. Właściwe przetwarzanie, w wyniku którego otrzymuje się konkordancje w jakiejś, być może pośredniej postaci, zwykle na nośniku komputerowym.

III. Stworzenie ostatecznej graficznie postaci wydruku konkordancji.

Należy przy tym wyraźnie powiedzieć, że przy podejmowaniu decyzji dotyczących danego etapu trzeba uwzględniać wpływ następnych etapów, ponieważ decyzje dotyczące np. trzeciego etapu mogą powodować istotne zmiany rozwiązań przyjętych dla etapu pierwszego czy drugiego.

Możliwie wczesne ustalenie, jakie oprogramowanie będzie użyte do przetwarzania i jakiego rodzaju informacje zawarte w tekście chcielibyśmy uwidocznić na ostatecznym wydruku, bardzo upraszcza dalszy tok postępowania.

Jeśli pominąć uwarunkowania finansowo-organizacyjne (będzie o nich mowa nieco dalej), to w wypadku oprogramowania możliwości było kilka. Po pierwsze, można było próbować napisać własny program lub programy, które umożliwiłyby wykonanie wszystkich potrzebnych prac, tzn. indeksów, konkordancji itd. Rozwiązanie takie daje możliwość uzyskania dokładnie takiego wyniku, jaki jest oczekiwany (pomijam tu warunki sprzętowe, np. dostępność odpowiedniej drukarki). Pisząc całe oprogramowanie od podstaw, można stosunkowo łatwo spełnić większość wymagań dotyczących postaci graficznej (wyglądu) np. konkordancji i jednocześnie uwzględnić specyfikę przetwarzanych tekstów.

Znacznie trudniej jest zapewnić efektywność takiego oprogramowania, a ponadto koszty jego tworzenia są na ogół bardzo wysokie. Przede wszystkim jednak praca tego rodzaju wymaga stałego dostępu do sprawnie działającego komputera i poświęcenia dużej ilości czasu przez dobrych programistów.

Druga możliwość to wykorzystanie istniejącego oprogramowania przeznaczonego do tego typu obliczeń. Chociaż systemów komputerowych przeznaczonych do tych celów jest bardzo dużo, w naszym zasięgu znalazły się tylko dwa z nich. Pierwszy to system COCOA, opracowany przez Atlas Computer Laboratory w latach 1973-74 i dostępny w IInf UW od 1975 roku dla komputerów ODRA 1300 i IBM360/370. Wcześniejsze doświadczenia z jego wykorzystaniem nie były zbyt zachęcające. W 1977 roku opracowano (por. Błażejczyk 1977), co prawda, na bazie COCOA program HASLO umożliwiający hasłowanie, ale był on bardzo niewygodny w użyciu, a co więcej, nie był dostępny na komputerze, na którym miały być wykonywane obliczenia. Tak więc w dostępnej dla nas wersji systemu brak było skutecznych mechanizmów wspomagających hasłowanie. Chociaż system COCOA wymaga dużej ilości czasu komputera, nie przewidziano w nim możliwości przechowywania pośrednich wyników obliczeń. Natomiast ze względu na błąd, którego nie udało się zlokalizować, nie działa mechanizm pozwalający na wykonywanie obliczeń przedziałami alfabetycznymi, którego użycie mogło się okazać w tej sytuacji konieczne. Na dodatek, kiedy podejmowana była decyzja wyboru oprogramowania, na maszynie IBM370, jedynej, która wchodziła w rachubę do naszych celów, system n ie był eksploatowany. Ogólnie rzecz biorąc, użycie COCOA nie wydawało się celowe.

Drugi system, którym dysponowaliśmy, to system LDVLIB opracowany przez R. Drewka na początku lat osiemdziesiątych (por. Drewek 1982, Drewek 1984) przewidziany do wykonywania bardzo różnych obliczeń statystycznych na tekstach (do zadań tego typu należy sporządzanie konkordancji). Projektując go, autor znał system COCOA, a więc można było mieć nadzieję, że LDVLIB będzie znacznie wygodniejszy i szybszy. Atrakcyjnym elementem systemu jest moduł wspomagający hasłowanie. LDVLIB jest wykorzystywany między innymi przez autora w projekcie konkordancji do "Operettte morali" G. Leopardiego (w języku włoskim, por. jak wyżej).

Podobnie jak inne systemy tego typu, LDVLIB może być eksploatowany wyłącznie na dużych komputerach. Autor opracował go dla dużych komputerów firmy IBM. W czasie pracy nad Konkordancjami jedynym dostępnym kom-

puterem tego typu był komputer IBM370/148 w Instytucie Organizacji Zarządzania i Doskonalenia Kadr (IOZiDK). W IInf podjęto około 18983 roku decyzję o sprowadzeniu systemu LDVLIB w celu zastąpienia nim przestarzałego systemu COCOA.

Istotnym czynnikiem mającym wpływ na wybór systemu były uwarunkowania organizacyjno-finansowe. System LDVLIB był narzędziem programistycznym, którego opanowanie leżało w kręgu zainteresowań IInf UW niezależnie od projektu Konkordancji. Planowane było zapoznanie się z tym systemem w ramach własnych prac badawczych i wykorzystywanie do tego celu opłacanego przez IInf UW czasu komputera IBM370/148 w IOZiDK. Możliwość przetestowania systemu LDVLIB na prawdziwym, dużym materiale dawała szansę obiektywnej oceny efektywności systemu LDVLIB.

Z drugiej strony budżet PSJCN nie pozwalał na finansowanie prac obliczeniowych prowadzonych w trybie zleceń czy zakupu czasu komputera.

Niestety otrzymana wersja systemu zawierała błąd, którego lokalizacja trwała wiele miesięcy, co spowodowało bardzo znaczne opóźnienie dalszych prac.

2.1. Wprowadzanie tekstu

Wprowadzanie tekstu jest chyba najważniejszym z trzech etapów wymienionych w poprzednim punkcie. Po pierwsze, w dużym stopniu decyduje on o uzyskanym efekcie. Po drugie, wymaga dużej ilości pracy ręcznej i powinien być tak przygotowany, aby nie trzeba było jej powtarzać. Można go podzielić na kilka kolejnych podetapów.

Pierwszy z nich, który omawiamy tu bardzo pobieżnie, to ustalenie tekstu kanonicznego. Za podstawę tekstu "Vade-mecum" zdecydowano się przyjąć tekst z wydania "Pism wszystkich" pod redakcją J. W. Gomulickiego (por. Norwid 1971). Sięgnięto jednak również do rękopisów. Okazało się, że teksty niektórych wierszy były w sposób istotny modyfikowane przez autora w procesie tworzenia. W niektórych przypadkach modyfikacje te prowadziły do powstania odrębnych rękopisów tego samego utworu. Ponieważ chodziło o przekazanie jak największej ilości informacji o tekście "Vade-mecum", początkowo planowano uwzględnienie tych modyfikacji poprzez wprowadzenie tzw. wariantów i odmianek. Zgodnie z przyjętymi przez PSJCN ustaleniami wariantem tekstu nazwano te fragmenty, które pochodziły z rozmaitych wersji rękopisu, a odmiankami — pojedyncze wyrazy, które autor po zapisaniu postanowił zmienić. Jeśli były to wyrazy inne niż użyte w tekście głównym, traktowano je jako odmianki leksykalne bądź słowotwórcze, natomiast nowe formy fleksyjne wyrazów użytych w tekście głównym traktowano jako odmianki fleksyjne. Ostatecznie w pierwszej wersji Konkordancji uwzględniono jedynie odmianki; natomiast wariantowe wersje niektórych wierszy, jako wymagające dalszego opracowania filologicznego, pozostawiono do późniejszego wprowadzenia (ewentualnie w ostatecznej wersji Konkordancji).

W wyniku tych ustaleń przygotowano tekst bazowy w postaci maszynopisu z naniesionymi odmiankami (będziemy jeszcze do nich wracać), będący podstawą dalszej pracy. Przy nieco innej organizacji pracy można było uniknąć użycia tradycyjnej maszyny do pisania. Można było z powodzeniem zastąpić ją komputerem. Dodatkową zaletą użycia komputera jest otrzymanie tekstu zapisanego na nośniku magnetycznym. Pozwala to uniknąć ponownego przepisywania całości.

Przetwarzanie tekstu przez program sterowane jest za pomocą odpowiednich instrukcji. Niestety bez wprowadzenia do przetwarzanego tekstu dodatkowych elementów niemożliwe jest przekazanie informacji o bilateralnych własnościach tekstu i o strukturze całego zbioru utworów.

W tym momencie należało już dość dokładnie wiedzieć, jak docelowo powinny wyglądać Konkordancje. Dokonany wcześniej wybór oprogramowania umożliwił natychmiastowe projektowanie sposobu realizacji ustaleń dotyczących Konkordancji. Niektóre realizacje wymagały wprowadzenia do tekstu pewnych dodatkowych informacji. Przykładem może tu być długość kontekstu na wydruku Konkordancji. Zasadniczo przyjęto, że powinien on zajmować cały wiersz wydruku (to znaczy — po odjęciu kilkunastu znaków informujących o lokalizacji hasła — około 120 znaków). Były jednak pewne ograniczenia, np. kontekst nie powinien przekraczać granicy strofy (z pewnymi wyjątkami), utworu, motta, tytułu itp. Z drugiej strony, przy takiej długości kontekstu ginęłyby granice pomiędzy wersami oryginału, co było sprzeczne z ogólnie przyjętą zasadą niegubienia informacji. Aby tego uniknąć, należało — przy wprowadzaniu tekstu — dodać oznaczenia końca wersu i specjalnie zaznaczyć miejsce, których kontekst absolutnie nie powinien przekraczać.

W systemie LDVLIB nie ma możliwości wprowadzenia specjalnych znaczników kontekstu (jak widać, autor nie przewidział konieczności tak skomplikowanych operacji na kontekstach). Kontekst może być natomiast ograniczony przez koniec zdania, to znaczy miejsce, gdzie po znaku alfabetycznym wystąpi bezpośrednio określony znak interpunkcyjny, np. kropka. Jeśli wprowadzi się do alfabetu dodatkowy znak (np. "/"), to po napotkaniu takiego znaku, a po nim kropki, system uzna to miejsce za nieprzekraczalną granicę kontekstu (sam znak "/" nie będzie taka granicą, może więc być użyty do oznaczania granic wersów). Spowoduje to jednak umieszczenie słowa "/" w indeksach i jako hasła w Konkordancjach. Aby tego uniknąć należy słowo "/" umieścić na tak zwanej liście wykluczeń (ang. stop-list, jest to specjalny mechanizm umożliwiający właśnie pominięcie słów jako haseł w konkordancjach, indeksach itp.). To, że kropki i inne znaki interpunkcyjne występujące w oryginalnym tekście nie będą powodowały zakończenia kontekstu, osiągnięto przez oddzielenie ich odstępem od słowa, po którym występowały.

Przyjmowane ostatecznie rozwiązania problemów pojawiających się w trakcie przygotowywania tekstu do wprowadzania były najczęściej kompromisem pomiędzy maksymalistycznymi oczekiwaniami przyszłych użytkowników Konkordancji (często natury czysto estetycznej) i względami (a właściwie ograniczeniami) technicznymi. Czynnikiem decydującym była zwykle zasada zachowania maksimum informacji o oryginale.

W dalszym ciągu niniejszego tekstu pojęciem wyraz określać będziemy jednostkę występującą w oryginalnym tekście Norwida, być może bilateralną — zawierającą elementy interpretacyjne i znaczeniowe. Natomiast terminu słowo używać będziemy technicznie — nazywając tak ciąg liter (czyli elementów alfabetu) zakończony odstępem lub znakiem niealfabetycznym, występujący w pewnej reprezentacji tekstu oryginalnego przetwarzanej przez komputer. Oprogramowanie przewidziane do obliczeń statystycznych (a takim jest system LDVLIB) traktuje przetwarzany tekst w sposób czysto formalny, zewnętrzny — jako obiekty unilateralne. W tej sytuacji podana wyżej definicja słowa jest właściwie definicją wymuszoną. Zapewnia ona sprawne przetwarzanie tekstu przez komputer i poprzez operowanie zestawem znaków alfabetycznych daje spore możliwości użytkownikowi. Istotne jest jednak, aby definicja ta była w pełni zrozumiała dla osób przygotowujących tekst do przetwarzania. W szczególności jej konsekwencją jest na przykład konieczność zdecydowania się, czy znaki wprowadzone do tekstu dla zaznaczenia jego cech, których inaczej nie można było przekazać, takie jak "-", "+", "@" itp. należą do alfabetu, czy też nie. Przykładem może być znak "@" użyty jako oznaczenie wyrazów podkreślonych przez autora. W Konkordancjach poprzedza on bezpośrednio (bez spacji) takie wyrazy. Wprowadzenie tego znaku do alfabetu spowodowałoby, że wyrazy nim poprzedzone pojawiłyby się w oddzielnej grupie jako hasła w konkordancji, nie wystąpiłyby natomiast pod literą, od której się zaczynają (np. słowo @czyny byłoby umieszczone pod 'literą' "@", nie zaś pod "c"). Z kolei niezaliczenie znaku do alfabetu pociągałoby za sobą niepożądany skutek: znak mógłby pojawić się na końcu cytatu sam, bez słowa następującego po nim, jeśli słowo to nie zmieściłoby się w obrębie linii wydruku. Zdecydowano się przyjąć to drugie rozwiązanie. Inną konsekwencją przyjętej definicji wyrazu jest fakt, że w Konkordancji nie da się potraktować wystąpień słów "Ojczyzna" i "ojczyzna" jako jednego hasła.

Należy jeszcze poświęcić kilka słów pojęciu alfabetu. Z pewnym uproszczeniem możemy założyć, że nazywamy tak ustalony zestaw znaków. Rzecz jasna w naszych rozważaniach występuje kilka alfabetów. Pierwszy — to alfabet używany przez Norwida. Różni się on znacznie od zestawu znaków, którym dysponuje maszyna. Jeszcze inne są alfabety (zestawy znaków), jakimi posługuje się program przetwarzający reprezentację tekstu. Operuje on trzema sprzężonymi alfabetami: wejściowym (w którym zapisany jest tekst do przetworzenia), alfabetem do sortowania (pozwalającym w pewnym stopniu kierować układem wyników) i wyjściowym (służącym do reprezentacji tekstu na wydruku). Używane obecnie w Polsce komputery (z wyjątkiem, być może, mikrokomputerów stosowanych do celów redakcyjnych) są słabo przystosowane do przetwarzania tekstów zapisywanych alfabetami innymi niż angielski (oprócz alfabetu rosyjskiego przy komputerach radzieckich). Dlatego zachodziła konieczność ścisłej transliteracji tekstu "Vade-mecum" na etapie wprowadzania tekstu. Zdecydowano się zastąpić litery opatrzone znakami diakrytycznymi — sekwencjami dwuznakowymi, np. litery z, ź, ż były wprowadzone jako z, z1, z2 (na wydruku natomiast uwidocznione jako z, 'z, "z). Wymagało to przewidzenia, jak te sekwencje będą się zachowywać w II etapie, w momencie sortowania (stąd właśnie cyfry 1 i 2), bowiem kolejność znaków z punktu widzenia komputera jest niezgodna z układem alfabetycznym.

Podczas opracowywania tekstu "Vade-mecum" przez PSJCN natknięto się na szereg problemów językowych, które należało rozwiązać. Poniżej przedstawiamy skrótowo opis tych problemów i przyjęte rozwiązania.

W toku dyskusji zdecydowano się rozdzielać i traktować jako złożone z samodzielnych jednostek wyrazowych stałe związki frazeologiczne oraz wyrażenia typu **po polsku**, **z cicha**, w których człon nominalny utracił samodzielność i stał się archaizmem fleksyjnym.

Poszczególne elementy analitycznych form czasownikowych, np. **będzie robił**, **robił był**, **(on) by (to) robił**, traktowano jako osobne wyrazy, zaopatrując je jednakże w odnośniki odsyłające wzajemnie do siebie, np.: **będzie+1 robił–1 był–1**, **on by+2 to robił–2**.

Odnośniki te wzorowane są na instrukcji do pracy (Woronczak 1983). Zastosowane tutaj wskazują, o ile wyrazów elementy analitycznej formy czasownika oddalone są od siebie. Przy liczeniu wyrazów od strony lewej do prawej liczby te poprzedzane są znakiem "+", przy liczeniu od strony prawej do lewej są poprzedzone znakiem "–".

Cząstkę **się** traktowano zawsze jako oddzielny wyraz.

Jako całość natomiast zdecydowano się traktować nazwiska poprzedzone cząstkami takimi jak **de**, **von**, np.: **de Musset**. Wymagało to odpowiedniego zakodowania tych fragmentów (o czym w odpowiednim momencie jednak zapomniano, skutkiem czego części **de**, **Musset** zostały potraktowane jako samodzielne wyrazy).

Kolejną konsekwencją przyjętej definicji słowa była konieczność rozdzielenia połączeń zawierających formy dwóch leksemów, np. połączeń z partykułami **-ż/że** (por. Saloni 1984), **-ć/-ci**, z formą zaimkową **-ń**, np. **dlań**, z ruchomymi końcówkami czasowników, np. **Jam widział**, a także połączeń z formami **-m, -ś, -śmy, -ście**, które można uznać za skrócone formy czasownika **być** w połączeniach typu **tyś pan, tyś król**. Dla zaznaczenia, że wymienione formy partykułowe, zaimkowe czy czasownikowe były pisane łącznie z innymi formami, w miejscu połączeń obu form stosowano znak "+", np.: **ty+ +ś**. W przypadku wystąpienia ruchomych końcówek czasownikowych stosowano przy nich i przy formach c czasownikowych, do których się odnosiły, odnośniki tego samego typu, co przy formach analitycznych, np. **Ja+ +m+1 widział–1** (Jam widział). Znak "+" zastosowano też dla skróconej formy przyimka **ku**, (tzn. formy **k'**, zastępując nim apostrof.

Charakterystyczną cechą pisarstwa Norwida jest łączenie za pomocą dywizu niezależnych (według naszej intuicji) jednostek leksykalnych (np. **ogień-święty**) bądź stosowanie tegoż dywizu w celu rozbicia wyrazu na rdzeń i szeroko rozumiany przedrostek, np. **u-wydatnić** (por. Subko 1986). Powstały wątpliwości, czy zapisy takie są jednym wyrazem, czy dwoma. Ostatecznie zdecydowano się traktować jako dwa wyrazy zapisy z dywizem pierwszego typu, o funkcji łączącej. Wyrazy tak zapisane powinny być reprezentowane przez dwa oddzielne słowa na liście alfabetycznej i być liczone oddzielnie.

Natomiast wyrazy z dywizami drugiego typu — o funkcji rozdzielającej — są reprezentowane jako jedno słowo na liście alfabetycznej. Dywizy nie zostały jednak usunięte (zgodnie z zasadą niegubienia informacji), ale zastąpione innym znakiem ("="), tak aby czytelnik zdawał sobie sprawę z użycia dywizu przez autora. Miało to również spowodować przetwarzanie słów je zawierających w odmienny sposób.

Wyrazy z dywizem o funkcji łączącej zostały zakodowane z użyciem znaku "-" (pauzy), bez dodatkowych spacji, czyli tak, jak występują w tekście oryginalnym. Niestety przyjęcie takiego sposobu kodowania spowodowało, że wyrazy z dywizem łączącym zostały potraktowane jako jedno hasło na liście alfabetycznej (znak "-" traktowany jest bowiem jako element alfabetu, czyli litera).

Znak "–" (minus poprzedzający cyfry, analogicznie jak "+") musiał być umieszczony w alfabecie, aby uwidocznić w wykazie haseł słowa stanowiące elementy większych całości, np. analitycznych form czasownikowych.

Niestety, na przyjętych przez nas rozwiązaniach dotyczących dywizów zaważył wyraźnie brak konsekwencji, który nie pozwolił na zastąpienie, w momencie kodowania tekstu, dywizów używanych przez Norwida, dwoma różnymi znakami (na wydruku konkordancji mogły one zostać zastąpione ponownie znakiem "-").

Drugą charakterystyczną cechą stylu Norwida jest podkreślanie ważnych wyrazów (co w wydaniach książkowych wyraża się poprzez użycie rozstrzelonego druku). Zdecydowano się wyrazy takie poprzedzić specjalnym oznaczeniem (użyto znaku "@", o którym była mowa wyżej).

Zasadnicza trudność przy opracowywaniu "Vade-mecum" stanowi niejednolitość formalna tego zbioru: zawiera on nie tylko wiersze, ale i fragmenty prozatorskie: wstępy, dedykacje, motta, przypisy. Znajdują się tu też fragmenty w językach obcych, zapisy cyfrowe, skróty. Każdy z wymienionych fragmentów tekstu wymagał odrębnego podejścia przy opracowywaniu. I tak, na przykład, o ile wiersze miały automatycznie numerowane wersy, o tyle fragmenty prozy były uważane za jeden wers (ze względu na brak wydania kanonicznego z numerowanymi wersami). Brak rozszerzenia tekstu głównego o specjalne dyrektywy dla programu LDVLIB spowodował jednak, że w lokalizacjach cytatów pojawiły się numery wersów prozy odnoszące się do ich układu w zbiorze wejściowym dla programu, a nie w tekście Norwida.

Spośród wyrazów czy fraz zapisanych w językach obcych uwzględniono w Konkordancji tylko te, które na podsta-

wie analizy zostały potraktowane jako integralna część wiersza, przypisu lub jako tytuł. Pominięto przypisy czy motta w postaci zdań zapisanych całkowicie w językach obcych, np.: **L'homme c'est le style**. Uwzględniono przekłady i trawestacje Norwida z utworów innych autorów.

Zapisy cyfrowe zostały pozostawione w cytatach, ale nie stanowią osobnych haseł.

Skróty zostały rozwiązane, z zachowaniem jednak informacji, że autor użył skrótu. Na przykład napis **C. N.** został wprowadzony jako **C[yprian] N[orwid]**.

Zdecydowano się umieścić w kontekstach odmianki. Odmianki leksykalne bądź słowotwórcze były były uwzględniane jako hasła przy przetwarzaniu tekstu. Natomiast odmianki fleksyjne miały zostać umieszczone w cytatach, ale nie powinny były stanowić osobnych haseł — pozycji w Konkordancji.

Zgodnie z przyjętą graficzną definicją słowa program LDVLIB nie umożliwia różnego traktowania poszczególnych słów, a tym samym wyrazów reprezentowanych identycznie. W szczególności odmianki fleksyjne również stanowią osobne hasła na liście alfabetycznej.

Przyjęto następujący sposób kodowania odmianek:

– wprowadzono dwa typy nawiasów w celu odróżnienia odmianek leksykalnych od fleksyjnych: są to nawiasy dwuznakowe postaci: /(i)/ dla odmianek leksykalnych oraz /[i]/ dla odmianek fleksyjnych,

– wewnątrz nawiasów umieszczane są dwie wersje tekstu, wersja główna oraz odmiankowa, oddzielone znakiem "/", przy czym jedna z wersji może być pusta, co oznacza skreślenie wyrazu lub dopisanie nowego, a nie zastąpienie wyrazu innym.

Przykład:

wers główny: **Począłże grać?**

wers odmiankowy: **Czy począł grać?**

są kodowane: **/(Począłże / Czy począł)/ grać ?**

Oprócz rozwiązania wymienionych wyżej trudności wiele czasu przeznaczono na dyskusję nad wyglądem Konkordancji. Najwięcej wątpliwości wzbudzała kwestia długości cytatu. Przyjęto rozwiązanie mechanicznego ograniczenia długości cytatu do jednej linii wydruku komputerowego, przy czym obowiązywała zasada umieszczania w cytacie tylko tych słów, które w całości mieściły się w obrębie linii wydruku. Cytat mógł się składać z fragmentów więcej niż jednego wersu, natomiast nie mogły się w nim w zasadzie znaleźć fragmenty więcej niż jednej strofy, (czy tytułu i tekstu wiersza, tekstu wiersza i przypisu itp.). Nie mogła, na przykład, stanowić jednego cytatu sekwencja zawierająca tytuł wiersza, dedykację i fragment tekstu wiersza: **FORTEPIAN SZOPENA. Do Antoniego C[zajkowskiego]. Byłem u Ciebie w te dni przedostatnie ...** Wymagało to zastosowania rozwiązań technicznych opisanych już wcześniej.

Jedynym wyjątkiem od tej zasady było łączenie sąsiednich strof wierszy, w sytuacjach uzasadnionych strukturą logiczną danego wiersza.

Samo wprowadzanie tekstu odbywało się w sposób dość skomplikowany organizacyjnie. Aby zrozumieć, dlaczego tak było, należy powiedzieć parę słów o uwarunkowaniach organizacyjnych, sprzętowych i finansowych.

Ze względu na wybrane oprogramowanie samo przetwarzanie (etap II) mogło się odbywać wyłącznie na dużym komputerze typu RIAD, IBM360 lub IBM370. Przestarzałe oprogramowanie systemowe (kompilatory), tryb pracy oraz zawodność wykluczały użycie sprzętu dostępnego na Uniwersytecie Warszawskim. Jedynym dostępnym komputerem odpowiadającym naszym potrzebom był IBM370/148 zainstalowany w IOZiDK, wykorzystywany przez IInf UW do własnych prac badawczych w wymiarze około 3 godzin tygodniowo. Chociaż Instytut był zainteresowany przetwarzaniem Konkordancji, nie widział potrzeby finansowania z własnych środków kosztownego procesu wprowadzania tekstu. Ponieważ PSJCN nie dysponowała funduszami na zakupienie czasu tego komputera, zdecydowano się na wykonanie tej pracy w Centrum Informatycznym IW (CIUW). Bardzo pożądana była możliwość interakcyjnego redagowania tekstu zawierającego wielkie i małe litery. Po przeanalizowaniu sytuacji CIUW zaproponował opisane niżej metody wykonania pracy.

Za pomocą mikrokomputera IMP 86 (odpowiednik IBM PC), tekst został zapisany na nośniku magnetycznym (dyskietka 5 1/4 cala w formacie IBM PC). Pracę tę wykonywał lingwista, będący jednocześnie informatykiem. Wydaje się, że mogła to robić osoba o skromniejszych kwalifikacjach, ale w tej sprawie nie udało się uzgodnić jednolitego stanowiska. Na mikrokomputerze IMP wykonano również pierwszą korektę wprowadzonych danych. Wszystkie te prace były wykonywane za pomocą niewygodnego w użyciu wierszowego edytora tekstowego edlin, udostępnionego przez CIUW razem z mikrokomputerem IMP.

Następnie tekst został przeniesiony na taśmę magnetyczną na minikomputerze SM-4 w CIUW (w formacie właściwym dla tego minikomputera).

Aby taśma ta mogła być użyta na IBM370, zawartość jej została przepisana zgodnie z formatem tej maszyny, na minikomputerze SM-4 w IInf UW, za pomocą opracowanego tam wcześniej oprogramowania.

2.2. Przetwarzanie

Samo przetwarzanie (etap II) odbyło się na komputerze IBM370/148 w IOZiDK. Komputer ten pracuje w trybie interakcyjnym (bezpośredniej komunikacji człowiek-maszyna za pomocą terminala ekranowego) pod systemem operacyjnym VM/CMS. Ponieważ system LDVLIB, zaprogramowany w języku PL/I, wymagał innej wersji kompilatora niż dostępna w systemie CMS, musiał być uruchomiony w systemie OS/VS1 dostępnym tam wsadowo (określa się tak wcześniejsze przygotowywanie zadań dla komputera na kartach perforowanych), z możliwością przygotowywania zadań w trybie interakcyjnym z terminali CMS.

Zanim przystąpiliśmy do przetwarzania całego tekstu "Vade-mecum" przeprowadziliśmy szereg prób, testując poszczególne przyjęte rozwiązania. Przeprowadziliśmy również próbne obliczenia na wybranym fragmencie, tak skomponowanym, aby zawierał trudniejsze miejsca występujące w tekście właściwym. Obliczenia te dały wyniki zadowalające. Niestety, jak się później okazało, próbka nie zawierała wszystkich typów fragmentów sprawiających kłopoty.

Aby przyśpieszyć pracę programu, dane wejściowe z taśmy zostały przekopiowane do zbioru dyskowego. Same obliczenia przebiegały dwuetapowo. W pierwszym etapie zapisana została taśma z wynikami pośrednimi i zbiorami roboczymi (zawierającymi między innymi indeks słowoform w postaci łatwo czytelnej dla programu). W drugim, korzystając z wyników zapisanych na taśmie w pierwszym etapie, wykonano indeks słowoform i Konkordancje w formacie typowym dla systemu LDVLIB. Zapisano je również na taśmie magnetycznej.

Zasadnicze obliczenia pochłonęły nieco ponad 12 minut czasu procesora (co odpowiadało mniej więcej dwóm godzinom czasu zegarowego), w tym pierwszy etap — ok. 6 minut, zaś wykonanie indeksu — ok. 1.5 minuty i Konkordancji — ok. 5 minut. Tym samym spełniły się nasze nadzieje, co do szybkości pracy systemu. Indeks słowoform składał się z 6060 opozycji. Wystąpień słów znajdujących się na stop-liście (a więc przede wszystkim "/") było około 2600.

2.3. Ostateczna postać Konkordancji

W wyniku drugiego etapu prac uzyskano Konkordancje zapisane na taśmie magnetycznej w formacie systemu LDVLIB. Największą wadą ich postaci było zastąpienie polskich liter ze znakami diakrytycznymi kombinacjami dwuznakowymi. Ponieważ Konkordancje liczą około 550 stron, ich wydrukowanie wymagało użycia szybkiej drukarki. Niestety, drukarka taka dysponująca ponadto małymi i wielkimi literami oraz literami z polskimi znakami diakrytycznymi nie była dostępna.

Postanowiliśmy wykorzystać drukarkę z wymiennym paskiem z małymi i wielkimi literami, którą dysponuje komputer BASF (funkcjonalny odpowiednik IBM370) znajdujący się w CIUW. Niestety, czego nie sprawdzono wcześnie, małe litery zostały wprowadzone przez producenta w taki sposób, że zestaw znaków dodatkowych został istotnie ograniczony i nieco zmieniony w stosunku do tego, jakiego się spodziewaliśmy. W szczególności nawiasy kwadratowe są tam zastąpione znakiem centa i wykrzyknikiem (wykrzyknik drukuje się jako pionowa kreska), zaś jeden ze znaków użyty przez nas do oznaczania diakrytów drukuje się jako spacja.

Ze względu na to, że narzucono nam krótki termin wykonania pracy, jedynym wyjściem było ponowne wykonanie całego (kosztownego) etapu II, przy nieznacznie zmodyfikowanym parametrze określającym alfabet wejściowy dla programu. Etap II został więc powtórzony.

Wydrukowanie jednego egzemplarza Konkordancji trwało około 40 minut i odbywało się w trybie nadzwyczajnym, ponieważ wymagany przez nas pasek nie był wówczas jeszcze włączony do eksploatacji w CIUW.

3. Podsumowanie i wnioski

Opisany przez nas proces istotnie okazał się sprawdzianem proponowanych metod przetwarzania tekstów C. Norwida i wykazał wyraźnie zarówno niedostatki organizacyjne i sprzętowe, które zmusiły nas do stosowania skomplikowanych

zabiegów, jak również wiele błędów czy niekonsekwencji w naszym postępowaniu. Uzyskany wynik można oceniać rozmaicie. Z puntu widzenia IInf UW, który finansował instalacje systemu LDVLIB i właściwe obliczenia, wynik jest pozytywny, ponieważ system został intensywnie przetestowany i zostały zbadane jego wymagania co do pamięci operacyjnej i pamięci dyskowej oraz czasu procesora przy dużych obliczeniach. Z punktu widzenia informatyków współpracujących z PSJCN wynik jest również pozytywny, ponieważ sporządzone konkordancje spełniają wszystkie te wymagania, które zostały przez Pracownię sformułowane w sposób pewny i jednoznaczny.

Z punktu widzenia PSJCN cel nie został jednak osiągnięty, ponieważ uzyskane konkordancje odbiegają wyglądem od pierwotnie oczekiwanych. Z pola widzenia pracowni umknął aspekt szkoleniowy całego przedsięwzięcia i użyteczność wniosków organizacyjnych dotyczących całej pracy. Mimo że uzyskany wynik nie jest w pełni zadowalający, Konkordancje okazały się użyteczne. Stały się one (łącznie z kartoteką Słownika) podstawą referatu (por. Engelking-Teleżyńska 1986). Ponadto służą one do weryfikowania kartoteki Słownika (jest to zastosowanie dyskusyjne) przygotowanej ręcznie. Można mieć nadzieję, że zostaną one jeszcze wykorzystane w dalszych badaniach prowadzonych przez PSJCN.

Na podstawie tego, co powiedziano dotychczas, można sformułować pewne wnioski.

Aby ewentualnie kontynuować prace lub doprowadzić Konkordancje do ostatecznej postaci, należy zweryfikować wszystkie podjęte dotychczas decyzje z zachowaniem formy pisemnej, usuwają oczywiście zauważone usterki koncepcji. W obecnej sytuacji prowadzenie dyskusji jest bardzo trudne, ponieważ istnieją tylko niektóre ostateczne pisemne wersje dokonywanych wcześniej ustaleń.

Ogólnie przy prowadzeniu prac tego typu na dużych tekstach należy mieć na uwadze następujące sprawy:

1. Komputer może być bardzo użytecznym przy wszelkiego rodzaju statystycznych pracach nad dużymi korpusami tekstów.

2. Bardzo istotnym elementem tego rodzaju prac jest zespół ludzi je wykonujących. Konieczne jest stworzenie grupy o jak najbardziej stabilnym składzie, złożonej z ludzi, dla których ta praca w danym momencie będzie najważniejszym (lub jedynym) zajęciem. Koordynacja prac powinna być prowadzona stale przez jedną osobę, umiejącą nie tylko sformułować jasno zapotrzebowanie końcowe, lecz również analizować i oceniać wyniki pośrednie.

3. Całe przedsięwzięcie musi być drobiazgowo zaplanowane zarówno merytorycznie, jak też organizacyjnie. Wybrany plan powinien zostać szczegółowo sprawdzony w praktyce na małej, ale w pełni reprezentacyjnej próbce.

4. Wszelkie projekty i ustalenia powinny być dokumentowane. Powstające teksty należy opatrywać datami i numerami wersji. Muszą one być zaakceptowane przez wszystkich zainteresowanych. W szczególności dotyczy to tekstu opisującego żądaną postać wyniku końcowego. Wszelkie najdrobniejsze nawet zmiany muszą być natychmiast nanoszone na teksty dotyczące ustaleń. Szczególną uwagę należy zwrócić na ewentualne sprzeczności, które mogą pojawić się w tych tekstach.

5. Aby uzyskać w pełni zadowalające konieczne są pewne minimalne warunki sprzętowe, przede wszystkim odpowiednie urządzenia wyjściowe i wejściowe oraz dostępność sprzętu komputerowego.

Literatura

[1] B. Błażejczyk **Modyfikacja systemu COCOA.** Praca magisterska. IInf UW, Warszawa 1977.

[2] R. Drewek, M. Erni **A system for interactive lemmatizing and its application.** COLING 82 Abstracts, Univerzita Karlova, Praha 1982, pp. 86-89.

[3] R. Drewek, M. Erni **LDVLIB: A (new) Software Package for Text Research.** ALLC Bulletin, vol.12, No.3(1984), pp 73-81.

[4] E. Engelking-Teleżyńska **Z badań nad strukturą ilościową "Vade-mecum".** Materiały z konferencji poświęconej problemom języka osobniczego. Zielona Góra, marzec 1986, red. J. Brzeziński (w druku).

[5] I. Kurcz, A. Lewicki, A.Sambor, J. Woronczak **Słownik współczesnego języka polskiego. Listy frekwencyjne.** Warszawa 1976.

[6] C. K. Norwid **Pisma wszystkie.** Zebrał, tekst ustalił, wstępem opatrzył Juliusz w. Gomulicki, Warszawa, 1971-76: Tom2: Wiersze, 1971, s. 457.

[7] J. Sambor **Słowa i liczby.** Wrocław, 1972.

[8] Z. Saloni **Instrukcja hasłowania dla SJCN**, 1984. [notatki maszynopisowe, 3 strony].

[9] B. Subko **O funkcjach łącznika w poezji C. Norwida.** [W:] "Studia Norwidiana" (w druku)

[10] (Woronczak 1983): **Jan Kochanowski, Dzieła wszystkie, Wydanie sejmowe:**
Tom wstępny: Wprowadzenie wydawnicze, Wrocław 1983; s. 48-53.
Tom I:Psałterz Dawidów. Część III Indeksy, Wrocław 1983.