Janusz S. Bień

Z PROBLEMÓW MASZYNOWEGO PRZETWARZANIA TEKSTÓW POLSKICH ZMODYFIKOWANA NOTACJA TOKARSKIEGO

Zainteresowanie Profesora Tokarskiego maszynami liczącymi jest powszechnie znane, choćby z racji cyklu artykułów publikowanych w „Poradniku Językowym" w latach 1961-1964 pod tytułem *Fleksja polska, jej opis w świetle możliwości mechanizacji w urządzeniu przekładowym*. Chciałbym się tutaj zająć niejako drugą stroną medalu, a mianowicie oddźwiękiem, jakie prace Tokarskiego znalazły w środowisku informatycznym. Skutkiem kontaktów Profesora Tokarskiego z informatykami były m.in. dwa artykuły, które zostały zamierzone przez autorów informatyków jako początek cyklu, opublikowane w roku 1970 i 1972 w „Poradniku Językowym" pod wspólnym nagłówkiem *Z problemów maszynowego przetwarzania tekstów polskich*. Ponieważ tekst niniejszy stanowi częściową odpowiedź na pytanie, jak potoczyły się losy zainicjowanych wówczas prac, uznałem za stosowne posłużyć się tym samym tytułem [1].

Pierwsze próby wykorzystywania komputerów do prac związanych z językiem polskim miały miejsce pod koniec lat sześćdziesiątych. W wypadku naszego zespołu inspiracja zajęcia się tą problematyką pochodziła od wówczas doktora, a obecnie docenta Stanisława Waligórskiego, prowadzącego wtedy seminarium z teorii maszyn na wydziale matematycznym Uniwersytetu Warszawskiego. Problematyka ta została podjęta przez kilku studentów; trzech z nich – Witold Łukaszewicz, Stanisław Szpakowicz i autor – zostało etatowymi pracownikami Uniwersytetu i po dziś dzień zajmuje się różnymi aspektami związków informatyki z lingwistyką. Założenia, które przyjęliśmy już na samym początku naszej pracy, można streścić następująco.

Użytkowe przetwarzanie tekstów języka polskiego nie jest możliwe bez rzetelnego rozwiązania problemów analizy i syntezy morfologicznej. Przez analizę rozumiemy przejście od danego ciągu liter (obecny stan techniki nie umożliwia jeszcze analizy tekstu mówionego) do możliwie pełnej informacji

[1] Pełną informację o obecnym stanie prac można znaleźć w pracach: J. S. Bień (ed.), *Papers in Computational Linguistic I*, „Sprawozdania Instytutu Informatyki" UW nr 107; J. S. Bień (ed.), *Paper in Computational Linguistics II*, „Sprawozdania Instytutu Informatyki" UW nr 110 oraz w literaturze w nich cytowanej.

o tym napisie jako formie fleksyjnej odpowiedniego wyrazu. Na przykład wynikiem analizy morfologicznej dla napisu *kaszy* będzie informacja, że jest to forma dopełniacza liczby pojedynczej rzeczownika, którego formą słownikową jest napis *KASZA*; analogicznie, wynik analizy napisu *straszy* to informacja, że jest to forma 3. osoby liczby pojedynczej czasu teraźniejszego czasownika, którego bezokolicznik ma postać *STRASZYĆ*. Synteza morfologiczna jest procesem odwrotnym – na podstawie odpowiedniej informacji o wyrazie tworzy się wskazane formy fleksyjne. Poprawna analiza i synteza morfologiczna wymaga odpowiednio obszernego słownika komputerowego, przechowującego informacje o poszczególnych wyrazach; inaczej napis *kaszy* mogły być potraktowane jako forma czasownika **KASZYĆ*, a napis *straszy* – jako forma rzeczownika **STRASZA*. Oczywiście, sama analiza morfologiczna nie potrafi rozpoznać, czy napis *kotka* jest mianownikiem rzeczownika *KOTKA* czy dopełniaczem lub biernikiem rzeczownika *KOTEK*; w takich wypadkach wynikiem jest lista możliwych interpretacji danego napisu [2].

Konsekwencją powyższych założeń było poszukiwanie źródła dostatecznie szczegółowej informacji o odmianie wyrazów. Szczęśliwie był nam znany *Słownik* PAN, pamiętaliśmy w szczególności o zawartym w nim stwierdzeniu, że „dobór wzorów w tabelach, odsyłaczy do nich i form podawanych przy haśle jest tak ułożony, by korzystający ze *Słownika* mógł odtworzyć pełny zasób form fleksyjnych danego wyrazu w zakresie uwzględnionym w *Słowniku*" [3]. Jednak już pierwszy, przeprowadzony w 1968 roku eksperyment pokazał, że opis tworzenia form fleksyjnych wymaga pewnych uściśleń [4]. Od tego momentu w naszym żargonie zaczęły funkcjonować dwa terminy: „notacja Tokarskiego", tj. przyhasłowa informacja gramatyczna w *Słowniku* PAN wraz z zasadami jej interpretacji, i „zmodyfikowana notacja Tokarskiego", tj. notacja Tokarskiego uzupełniona o uściślenia niezbędne dla właściwej jej interpretacji przez komputer.

Pierwszy wariant zmodyfikowanej notacji Tokarskiego dla czasowników powstał w roku akademickim 1968/69. Ze względów technicznych informacja przyhasłowa miała w niej inną formę zewnętrzną, ale jedyne istotne zmiany w stosunku do oryginalnej notacji polegały na pełnym sprecyzowaniu zasad

---

[2] Stosując terminologię pracy J. S. Bienia i Z. Saloniego *Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna)*, „Prace Filologiczne" XXXI (1982), s. 31-45, można określić analizę morfologiczną jako odszukanie tych wyrazów, których kształt grafemiczny jest zgodny z danym napisem (z dokładnością do pewnych cech typograficznych, takich jak duża litera na początku zdania, przeniesienie słowa do nowego wiersza itp.); syntezę morfologiczną można opisać jako utworzenie odpowiednich kształtów grafemicznych, gdy znany jest reprezentant paradygmatyczny (forma słownikowa) wyrazu i wartości właściwych dla danej części mowy kategorii fleksyjnych.

[3] Jan Tokarski, *Formy fleksyjne*, (w:) SJPD t. I, s. L.

[4] Por. J. S. Bień, *Algorytmizacja fleksji polskiej – problemy i perspektywy*, „Maszyny Matematyczne" 1969, nr 5, s. 15-18.

interpretacji. Eksperymenty przeprowadzone z tą wersją notacji [5] potwierdziły nasze wcześniejsze obawy, że przygotowana przez profesora Tokarskiego instrukcja dotycząca informacji fleksyjnej w artykułach hasłowych nie była przez redaktorów poszczególnych haseł dostatecznie konsekwentnie stosowana. Stanęliśmy więc przed zadaniem przerastającym nasze siły, a mianowicie przed zadaniem weryfikacji informacji przyhasłowej dla wszystkich wyrazów wprowadzonych do komputerowego słownika. Nie poddaliśmy się jednak od razu, jeszcze w 1971 r. została wykonana praca magisterska pod tytułem „Badanie struktury słownikowej informacji czasownikowej dla potrzeb systemu konwersacyjnego". Dokonana przez Marię Franjasz analiza dotyczyła tylko formy informacji przyczasownikowej, ale jej efektem ubocznym było około 17 000 fiszek zawierających wszystkie czasowniki ze *Słownika* PAN wraz z informacją przyhasłową. Ciekawostką jest fakt, że fiszki te zostały sporządzone z własnej inicjatywy magistrantki i to praktycznie bez wiedzy opiekuna pracy.

Zadanie zweryfikowania informacji gramatycznej zostało podjęte na nowo dopiero w roku 1978, tym razem przez docenta Zygmunta Saloniego. Oprócz kartoteki mgr Franjasz przejął on około 55 000 fiszek z rzeczownikami, sporządzonych na Uniwersytecie Poznańskim pod kierunkiem wówczas docenta, a obecnie profesora Zygmunta Zagórskiego. Jednocześnie sformułował on zestaw tematów prac magisterskich, które — zrealizowane przez studentki Filii UW w Białymstoku [6] — doprowadziły do utworzenia zbiorczej kartoteki zawierającej wszystkie hasła *Słownika* PAN łącznie z suplementem. Informacja przyhasłowa została przy tym sprawdzona i w razie potrzeby skorygowana — poprawiono błędy drukarskie, a także pomyłki redaktorów słownika. Dla haseł rzeczownikowych zmodyfikowano informację rodzajową, rozróżniając trzy rodzaje męskie. Informacja gramatyczna zawarta w kartotece stanowi aktualną, choć być może jeszcze nie ostateczną wersję zmodyfikowanej notacji Tokarskiego.

Tak więc w ponad 100 lat od sformułowania przez Karłowicza postulatu, aby słowniki podawały sposób odmiany wyrazów, w około 30 lat od sformułowania przez profesora Tokarskiego koncepcji realizacji tego postulatu, w około 15 lat od pierwszych prób wykorzystnia tej koncepcji dla potrzeb informatyki, dysponujemy wreszcie obszernym i wiarygodnym opisem morfologicznym polskiego słownictwa. Po wprowadzeniu go do komputera może on znaleźć różnorodne zastosowania — niektóre z nich mogą być użyteczne

[5] Por. prace magisterskie zrealizowane w Instytucie Maszyn Matematycznych UW w 1970 r.: W. Król, Transformacja notacji słownikowej na zmodyfikowaną notację Tokarskiego — program na maszynę GIER; S. Szpakowicz, Optymalizacja reprezentacji leksemów czasownikowych w maszynie.

[6] Annę Barszczewską, Halinę Jastrzębską, Danutę Kaniecką, Jadwigę Kruczewską, Marię Kulikowską, Halinę Lipińską, Leontynę Naruszewicz, Danielę Nowacką, Joannę Raszkowską, Joannę Saniewską, Henrykę Wasilewicz.

natychmiast, inne pojawią się w miarę tego, jak polska informatyka zacznie się zbliżać do obecnego poziomu światowego.

Przykładem pierwszej grupy zastosowań może być sporządzanie wszelkiego rodzaju indeksów, które wymagają przekształcenia konkretnych form fleksyjnych na postaci słownikowe odpowiednich wyrazów; dotyczy to zarówno indeksów przeznaczonych bezpośrednio dla człowieka (np. indeksy w książkach i innych wydawnictwach), jak i indeksów wykorzystywanych w komputerowych systemach wyszukiwania informacji. Druga grupa zastosowań, to np. procesory tekstowe (ang. *word processors*), czyli mikrokomputery wypierające stopniowo klasyczne maszyny do pisania; coraz częściej są one wyposażone w programy wykrywające błędy ortograficzne przez porównywanie słów tekstu z odpowiednim słownikiem (na razie dla języka angielskiego). Niezależnie od możliwych zastosowań praktycznych, zrealizowanie komputerowego słownika morfologicznego umożliwi pogłębienie wiedzy o morfologii polskiej, a dysponowanie programami analizy i syntezy morfologicznej ułatwi badanie wyższych pięter języka polskiego.

Reasumując, niezależnie od tego, jakie będą losy oryginalnej notacji Tokarskiego w polskiej praktyce leksykograficznej, zmodyfikowana notacja Tokarskiego ma zapewnioną przyszłość w informatyce, gdyż odpowiada ona na konkretne, choć nie w pełni jeszcze rozbudzone zapotrzebowanie, nie mając przy tym żadnej konkurencji.