## Korekta poprawności ortograficznej w dokumentach; stan obecny i perspektywy rozwoju

Włodzimierz Macewicz

*W chwili obecnej większość dokumentów jest tworzonych za pomocą komputera. W związku z tym naturalnym odruchem jest chęć skłonienia go do sprawdzenia poprawności ortograficznej wprowadzonego tekstu (być może także do kontroli składni).*

*Kontrola poprawności ortograficznej w języku polskim jest trudna do zautomatyzowania; posiada on zbyt wiele różnych odmian (i mnóstwo pojedynczych odstępstw). Znane mi słowniki „algorytmiczne" nie zauważały użycia wyrazów dziwnych – nie używanych w mowie – tworzonych przez algorytm danej odmiany (np: „dzieciniono"). Poprawność ortograficzna zależy również od kontekstu (np. „może" i „morze"). W chwili obecnej wszystkie znane mi programy sprawdzają poprawność wyłącznie wg zasady: słowo znajdujące się w słowniku jest uznawane za poprawne.*

*Słowa w słowniku mogą być przechowywane w sposób jawny (lista słów i ich odmian) lub w postaci spakowanej (rdzeń oraz kwalifikator mówiący jak generować formy dla danego rdzenia). W pierwszym wypadku pełen słownik języka polskiego zajmowałby ok. 12 MB (ok. 1.5 miliona słów), a w drugim ok 3 MB.*

*Do rozwiązania problemu kontroli ortografii potrzebne są dwie rzeczy:*

- *program sprawdzający*
- *słownik (zbiór słów danego języka).*

### Programy

Jest wiele wordprocesorów[1], które udostępniają funkcję sprawdzającą. Najbardziej ergonomicznie jest to zrobione w `MS Wordzie` – po wprowadzeniu wyrazu jest on natychmiast weryfikowany i w przypadku braku jego definicji w słowniku jest on podkreślany czerwoną linią. Pozostałe programy tego typu: `WordPerfect`, `AmiPro` mają również możliwość kontroli poprawności ortograficznej.

Podstawową wadą wyżej wymienionych programów jest fakt, że należy je kupić; inne wady to:

- dostępność wyłącznie dla systemu operacyjnego `MS Windows`[2];
- wielkość pakietów (potrzebują dużych zasobów pamięci oraz szybkiego procesora);
- zawodność programów;
- niestabilność formatu dokumentów.

Możemy postawić śmiałą tezę, że dla osób przygotowujących dokumenty składane TEX-em, wyżej wymienione programy są bezużyteczne (inny format dokumentu, wyłącznie system operacyjny `MS Windows`). Mając na uwadze wielką liczbę **absolutnie zbędnych** funkcji – za które niestety musimy zapłacić – zakup jest ekonomicznie nieuzasadniony.

Skoro odrzuciliśmy produkty czołowych firm softwarowych to narzuca się pytanie, co mamy innego do dyspozycji? Najpierw uspokoję nastroje wśród użytkowników systemu `MS Windows` – jest dobry, w pełni konfigurowalny edytor `WinEDT`[3]. Edytor ten został stworzony specjalnie dla TEX-owców i posiada wiele funkcji ułatwiających pisanie dokumentów składanych TEX-em. Oprócz tego ma możliwość podłączenia własnego słownika i podobnie jak `MS Word` podkreśla na czerwono wyrazy, których nie ma w słowniku. W dalszej kolejności mógłbym polecić program `WinPad`; zapisuje on dokument w „czystej" postaci oraz ma możliwość uruchomienia funkcji kontroli poprawności – niestety w przypadku dużych słowników (rzędu 400 tys. słów) trwa to długo.

Dla użytkowników innych systemów (w tym `MS-DOS`) mam kilka propozycji bazujących na dwóch programach kontroli poprawności: `ispell` oraz `texspell` (wraz ze swoim protoplastą `texortho`[4]).

Program `texspell` może pracować jedynie w trybie „off-line" tzn. możemy poddać kontroli jedynie gotowy dokument (w czasie innym niż chwila jego tworzenia); w wyniku kontroli otrzymujemy listę błędów z zaznaczeniem miejsca ich występowania. Używając programowalnego edytora możemy wykorzystać wygenerowaną listę błędów do sprawnego poprawienia dokumentu. Program używa słownika

1: takiego słowa w moim słowniku nie ma ;-)

2: `WordPerfect` dostępny jest dla `MS-DOS-a` i paru implementacji `Unix-a`, a `MS-Word` dla `Mac-a`

3: Znajduje się on na `CD-ROMie TEX Live 3` rozprowadzanym przez `GUST`

4: `texspell` to `texortho` z gruntownie przerobioną obsługą drzew binarnych (na których oparty jest proces poszukiwania słowa w słowniku) oraz poprawionym algorytmem uwzględniającym składnię poleceń TEX-a zapisanego w sposób jawny – lista słów wraz z odmianami. Wadą takiego rozwiązania jest długość słownika. Zasób słów równoważny popularnemu, podręcznemu słownikowi to ok. 3 MB pamięci dyskowej. Dla krótkich dokumentów (do ok. 100 KB – znacznie mniejszych niż długość słownika), a takie zdarzają się najczęściej, czas kontroli jest proporcjonalny do długości słownika, a nie długości dokumentu, co daje nieprzyjemne wrażenie wolnego działania programu.

Podaną wielkość słownika (3 MB) należy traktować jako maksymalną dla tego programu – wg mojej oceny, zapewnia to jeszcze minimum niezbędnego komfortu pracy na komputerze PC 486 66 MHz; jest to jednak zasób słów (ok. 300 tys.) w pełni wystarczający – nie musi zawierać on rzadziej używanych odmian słów.

Program `ispell` wydaje mi się najciekawszym rozwiązaniem. Może on pracować zarówno w trybie „off-line", jak i interakcyjnym pod kontrolą praktycznie dowolnego systemu operacyjnego. Pierwszy tryb doskonale nadaje się do współpracy z edytorami programowalnymi (np. Emacs), natomiast drugi umożliwia korektę dokumentu bez udziału dodatkowych programów. Jedynym zastrzeżeniem do tego trybu pracy jest konieczność wprowadzenia całego poprawnego słowa, zamiast błędnego – brak możliwości poprawienia pojedynczej literówki. Prace nad tym mankamentem trwają; jest już gotowa zmodyfikowana wersja dla systemu `MS-DOS`[5].

Słownik dla tego programu ma specjalną budowę (oczywiście razem z programem `ispell` są dostarczane programy do jego budowy). Są w nim przechowywane rdzenie (część główna słowa w sensie wyrażeń regularnych) oraz kwalifikator (flaga) określająca jakie odmiany mają być automatycznie wygenerowane. Zaletą takiego rozwiązania jest znacznie mniejsza objętość słownika, a co za tym idzie czas dostępu. Zaimplementowany w programie algorytm tablic mieszanych (*hashowanych*) daje lepsze efekty niż drzewa binarne, w dodatku program nie wczytuje do pamięci całego słownika, tak że czas sprawdzania dokumentu jest prawie proporcjonalny do jego długości. Warto podkreślić, że program proponuje listę słów podobnych do nieznanego, tak że w większości wypadków wystarczy podać numer właściwej propozycji.

5: Przy okazji został usunięty błąd w obsłudze poleceń TEX-a typu: `\,` `\#` itp

## Słownik

Są dwa sposoby przechowywania zasobu słów: jako listę słów i ich odmian oraz listy rdzeni i kwalifikatora oraz przepisu na generowanie odmian dla danego rdzenia wg oznaczonego kwalifikatora (są oczywiście różne implementacje tego sposobu). W pierwszym przypadku możemy przechowywać jedynie najczęściej używane słowa, rozszerzanie zasobu słów jest proste i polega na dopisaniu do słownika nowego słowa. Wadą jest szybki wzrost jego wielkości, a zatem czas kontroli dokumentu.

W programie `ispell` zastosowano drugą metodę, gwarantującą szybki czas dostępu. Wadą tego rozwiązania jest kłopotliwy sposób budowy słownika. Po pierwsze należy zdefiniować reguły odmian różnych części mowy, mając na uwadze wszystkie możliwe sytuacje i przełożyć je na język wyrażeń regularnych. Prace nad tym zagadnieniem są na ukończeniu (autorem reguł jest Piotr Gackiewicz z AGH). Drugim etapem jest zakwalifikowanie rdzeni (form podstawowych słów np.: bezokolicznik, rzeczownik w mianowniku liczby pojedynczej itp.) do właściwej reguły generacji odmian (opatrzenie ich właściwym kwalifikatorem). Prace nad tym etapem trwają. Zadanie pierwotne zostało uproszczone; z dostępnego zasobu słów zostały wybrane rdzenie i automatycznie opatrzone najbardziej prawdopodobnym kwalifikatorem. Zachęcam wszystkich potencjalnych użytkowników słownika do wzięcia udziału w pracach (wymiana informacji odbywa się na forum listy `slownik@listserv.ia.pw.edu.pl`); polega ona na zweryfikowaniu przydzielonego kwalifikatora (sprawdzeniu czy wszystkie wygenerowane odmiany są poprawne).

Zawiązał się zespół koordynujący prace nad słownikiem. Docelowo zasób słów ma być podzielony na kilka kategorii: słowa potoczne, używane rzadziej, podobne do częstych literówek (np. „swoich" i „sowich"; słowo „sowie" powinno znaleźć się w trzeciej grupie) oraz wiele zbiorów zawierających słowa branżowe (np. związane z medycyną, matematyką itp.), nazwiska, skróty. Użytkownik będzie mógł zbudować sobie słownik zawierający interesujący go zestaw słów. Ze względu na ogrom prac (ok. 150 tys. form podstawowych i ok. 2 mln wszystkich dostępnych słów) zespół zajmujący się tworzeniem słownika prosi o pomoc wszystkich chętnych.

◇ Włodzimierz Macewicz
`Macewicz@ia.pw.edu.pl`