Wybrane standardy przetwarzania tekstów*

dr hab. Janusz S. Bień
Instytut Informatyki, Uniwersytet Warszawski
Banacha 2, 02–097 Warszawa
JSBIEN@PLEARN.edu.pl

15 listopada 1993

1 Wstęp

W niniejszym artykule chciałbym zwrócić uwagę na dwa aspekty krajowej i międzynarodowej działalności normalizacyjnej związanej z przetwarzaniem tekstów. Po pierwsze, istnieją już gotowe normy regulujące różne zagadnienia związane z językiem naturalnym, które można wykorzystać w ich oryginalnej postaci lub jako punkt wyjścia do własnych rozwiązań, unikając w ten sposób wyważania otwartych drzwi i zaczynania od zera. Po drugie, normy takie są wynikiem pracy zespołów mniej lub bardziej kompetentnych osób o różnej motywacji i podległych różnym naciskom, dlatego też warunkiem dobrej jakości ustanawianych norm jest permanentna kontrola społeczna w formie rzeczowej dyskusji norm już w fazie ich opracowywania, która powinna prowadzić do eliminowania niekompetentnych autorów, tłumaczy, opiniodawców i weryfikatorów oraz wpływać na udoskonalanie organizacji prac normalizacyjnych i lepsze wykorzystanie dostępnych środków.

2 Polska klawiatura komputerowa

Użytkownicy komputerów osobistych nie odczuwają specjalnie braku jednolitej konwencji wprowadzania polskich znaków z klawiatury komputera, ponieważ z reguły mogą one być łatwo zmieniane w razie potrzeby; osoby piszące metodą bezwzrokową korzystają przeważnie z tzw. klawiatury maszynistki czyli układu klawiszy maksymalnie zbliżonego do maszyny do pisania, inni użytkownicy natomiast zadawalają się tzw. klawiaturą programisty czyli konwencję wprowadzania polskich liter za pomocą dodatkowego klawisza modyfikującego funkcję

*Referat wygłoszony na konferencji **Komputerowa Analiza Tekstu**, Karpacz, 16–18 listopada 1993, zorganizowanej przez Instytut Filologii Polskiej Uniwersytetu Wrocławskiego i Seminar für Slavistik, Universität Bochum

klawiszy literowych. Jednak przy zdalnej pracy na komputerach wielodostępnych korzysta się często z terminali pracujących w tzw. trybie echa lokalnego, kiedy znaki odpowiadające naciśniętym klawiszom są wyświetlane na ekranie przez terminal, bez angażowania w ten proces samego komputera; choć dopuszczenie różnych konwencji wprowadzania polskich liter jest technicznie możliwe, komplikuje to jego budowę i podwyższa cenę. Istnienie różnych grup użytkowników przyzwyczajonych do różnych konwencji jest także kłopotliwe dla firm produkujących oprogramowanie na rynek polski — tylko największe z nich mogą sobie pozwolić na lansowanie swoich rozwiązań jako jedynie słusznych, inne mają dodatkowe zajęcie starając się uwzględnić co najmniej dwie wspomniane wyżej konwencje, tj. klawiatury maszynistki i programisty.

Pierwsza dyskusja dotycząca normalizacji polskiej klawiatury komputerowej odbyła się 18.XI.1991 podczas Pierwszego Forum Technologii Informatycznych zorganizowanego przez Polskie Towarzystwo Informatyczne. Ustalono wówczas pewne podstawowe założenia, które zostały później uszczegółowione na spotkaniu 28.II 1992, na którym Piotr Carlson, wówczas pracownik firmy UNILOT (reprezentującej UNISYS), przedstawił propozycję modyfikacji normy PN-87/F-02000 (por. np. [3]) tak, aby mogła ona łączyć funkcję klawiatury maszynistki (podstawowy układ klawiszy — QWERTZ — przeniesiony z maszyny do pisania) i programisty (dostępność wszystkich znaków niezbędnych przy programowaniu). W zebraniu wzieli udział przedstawiciele kilkunastu firm komputerowych, a także Krzysztof Gujski, kierownik Zakładu Normalizacji i Badania Jakości w Instytucie Maszyn Matematycznych, oraz przedstawiciele PKNMiJ. Istotną trudnością w nadaniu przyjętym ustaleniom formalnego charakteru polskiej normy była odbywająca się właśnie reorganizacja działalności normalizacyjnej w Polsce. Dlatego też drugie zebranie — zorganizowane również przez Piotra Carlsona, ale tym razem jako pracownika firmy Digital Equipment Polska — odbyło się 2.X.1992, już po rozpoczęciu organizacji Normalizacyjnej Komisji Problemowej do spraw Informatyki przy Zespole Elektryki Polskiego Komitetu Normalizacji, Miar i Jakości. Na zebraniu zapoznano się z opinią Jana Woszczatyńskiego, pełniącego w Stowarzyszeniu Stenografów i Maszynistek funkcję Przewodniczącego Komisji do spraw Nauczania; opinia stwierdzała jednoznacznie, że należy przyjąc anglosaski układ klawiatury QWERTY, wprowadzono więc stosowne zmiany i sformułowano wniosek do Prezesa PKNMiJ o ustanowienie Polskiej Normy na klawiaturę komputerową. Sprawa klawiatury była dyskutowana kilka dni później na zebraniu założycielskim NKPI w dniu 15.X.1992.

Opracowanie tekstu normy PKNMiJ powierzył jednej z pracownic Zakładu Normalizacji IMM; z perspektywy czasu widać, że co najmniej współautorem normy powinien być ktoś z jej inicjatorów. Projekt normy PN- /T-42117 został rozesłany do tzw. ankietowania wybranym instytucjom i osobom, a do szerszych kręgów użytkowników starano się dotrzeć za pośrednictwem prasy komputerowej (jednak notka w *ComputerWorld PL* ukazała się dopiero 21.XII.1992 w nr 44/77); dnia 7.XII.1992 odbyła się w Instytucie Maszyn Matematycznych tzw. konferencja uzgadniająca. Projekt normy został przygotowany niestarannie i niekompetentnie (za co tylko częściowo można winić jego autorkę), ale konferencję zdominował spór merytoryczny — układ QWERTZ jak w wersji z

28.II.1992, czy układ QWERTY jak w wersji z 2.X.1992? Tomasz Lesz przedstawił jako oficjalne stanowisko Zarządu Stowarzyszenia Stenografów i Maszynistek pogląd odmienny od stanowiska Przewodniczącego Komisji do spraw Nauczania tegoż stowarzyszenia, uznający jako jedyny dopuszczalny układ klawiatury układ QWERTZ. Merytoryczną dyskusję praktycznie uniemożliwiły czasochłonne spory formalne i kompetencyjne przewodniczącego konferencji Krzysztofa Gujskiego z Tomaszem Leszem i przedstawicielem firmy PREBOT, również przeciwnym przedstawionemu projektowi. Będąc obecnym na zebraniu z ramienia Instytutu Informatyki Uniwersytetu Warszawskiego, powoływałem się na stanowisko Kotarbińskiego, że jeśli eksperci (w tym wypadku ze Stowarzyszenia Stenografów i Maszynistek) nie są zgodni, uprawnione jest podjęcie decyzji arbitralnej. W głosowaniu przedstawiony projekt przeszedł większością jednego głosu (5 za, 4 przeciw, 3 wstrzymujące się), co z czysto formalnego punktu widzenia było wystarczające do ustanowienia normy, zwłaszcza że opinie pisemne były raczej pozytywne. PKNMiJ wolał jednak przekazać odpowiedzialność za tę decyzję rozpoczynającej pracę z dniem 1.I.1993 Normalizacyjnej Komisji Problemowej do spraw Informatyki.

Przewodniczący NKPI Stanisław Koślacz zwołał specjalne spotkanie w tej sprawie na dzień 5.IV.1993; w jego wyniku na posiedzeniu NKPI dnia 21.IV.1993. przedstawiono koncepcję kompromisową dopuszczającą jako równoprawne warianty QWERTY i QWERTZ. Wszystkie osoby zainteresowane merytorycznie sprawą klawiatury zaakceptowały ten kompromis, natomiast nieoczekiwany sprzeciw zgłosili zawodowi normalizatorzy, domagając się — bez wskazania konkretnych przepisów czy wytycznych — wyróżnienia jednego wariantu jako preferowanego. Z innych pozycji przeciwko projektowi wystąpił Andrzej Gecow, który w swojej notatce napisał: *Wydaje się konieczne skokowe przejście do całkiem nowej polskiej klawiatury, opartej na innych założeniach niż dotychczasowa.* Tym niemniej postanowiono przygotować nowy projekt i przedyskutować go na następnym posiedzeniu 26.V.1993, które jednak nie przyniosło rozstrzygnięcia ze względu na brak quorum. Uzupełniające głosowanie korespondencyjne również nie pozwoliło osiągnąć quorum, ale gdyby nie stanowsko Andrzeja Gecowa byłaby możliwość osiągnięcia consensusu dzięki wycofaniu sprzeciwów formalnych.

Przez cały czas dyskusji nad polską klawiaturą komputerową przewijała się kwestia jej stosunku do znajdującej się w opracowaniu normy międzynarodowej ISO/IEC 9995 *Information technology — Keyboard layouts for text and office systems*, omówionej w punkcie następnym. Sprawą tą miała się zająć powołana z mojej inicjatywy na posiedzeniu NKPI dnia 29.VI.1993 Grupa Zadaniowa do spraw klawiatur, której przewodniczącym został Andrzej Gecow. Kiedy po przerwie wakacyjnej próbował on zorganizować zebranie tej grupy, okazało się to jednak niemożliwe ze względu na brak czasu zainteresowanych osób. Z tego samego powodu nie udało się dotąd — o ile mi wiadomo — uzyskać żadnych uwag ani komentarzy do opracowanego przezeń *Raportu o standardach klawiatury polskiej* (wersja 1 jest datowana na 10.X.1993) i nowej propozycji układu klawiatury; co więcej, w związku z moją rezygnacją z członkostwa w NKPI, grupa do spraw klawiatury zredukowała się do jej przewodniczącego.

Tak wyglądają pokrótce główne wątki burzliwych losów polskiej normy na klawiaturę komputerową; pominąłem m.in. historie różnych zmian redakcyjnych i merytorycznych. O ile początkowo byłem zdecydowanym zwolennikiem ustanowienia normy w wersji z 2.X.1992 lub 26.V.1993, malejące zainteresowanie sprawą ze strony przedstawicieli przemysłu komputerowego oraz coraz większe wątpliwości, czy ustalenia tych projektów są zgodne z duchem i literą ISO/IEC 9995, sprawiają łącznie, że nie mam obecnie w tej kwestii wyrobionego zdania.

3 Klawiatura międzynarodowa

Norma międzynarodowa ISO/IEC 9995 *Information technology — Keyboard layouts for text and office systems* (*Disposition des claviers conçus pour la bureautique*) jest wynikiem pracy grupy roboczej WG 9 podkomitetu SC 18 (*Document Processing and Related Communication*) wspólnego komitetu technicznego JTC1 (*Joint Technical Committee*) ISO (*International Standard Organisation*) i IEC (*International Electrotechnical Commission*). Polska jest członkiem czynnym JTC1, ale do niedawna nie była członkiem SC 18. Oznacza to praktycznie, że nie docierały dotąd do kraju materiały robocze SC 18, lecz tylko wyniki końcowe podlegające głosowaniu na posiedzeniach JTC 1. W konsekwencji długa i skomplikowana historia normy ISO/IEC 9995 nie jest u nas znana.

Aktualnie norma ta jest już zatwierdzona jako standard międzynarodowy pod względem merytorycznym, ale prace redakcyjne zakończyły się dopiero kilka tygodni temu; ostateczna postać tekstu normy powinna nadejść do PKNMiJ lada moment, ale w momencie pisania tych słów dysponuję jedynie projektem normy (DIS czyli *Draft Internationa Standard*) z lipca 1991 roku oraz nieoficjalnymi informacjami o wprowadzonych do tego projektu obszernych zmianach. Tak więc ograniczę się tutaj do podania tylko podstawowych ustaleń normy, zachęcając zainteresowanych czytelników do zapoznania się z oryginalnym tekstem, gdy tylko to będzie możliwe.

Norma ISO/IEC 9995 ma na celu uporządkowanie i ujednolicenie postaci klawiatur nie tylko w takich urządzeniach, jak maszyny do pisania, komputery osobiste, terminale itp., ale również w kalkulatorach, telefonach klawiszowych i bankomatach; w konsekwencji unieważnia ona kilkanaście wcześniejszych norm dotyczących klawiatur. Norma składa się z następujących części:

Part 1: *General principles governing keyboard layouts.* Ogólne zasady układu klawiatury.

Part 2: *Alphanumeric section.* Sekcja alfanumeryczna klawiatury.

Part 3: *Common secondary layout of the alphanumeric zone of the alphanumeric section.* Wspólny dodatkowy układ klawiatury w strefie alfanumerycznej sekcji alfanumerycznej.

Part 4: *Numeric section.* Sekcja numeryczna klawiatury.

Part 5: *Editing section.* Sekcja edycyjna klawiatury.

Part 6: *Function section.* Sekcja funkcyjna klawiatury.

Part 7: *Symbols used to represent functions* (*Symboles employés pour la representations des fonctions*). Symboliczne oznaczenia klawiszy funkcyjnych; ta część normy jest dwujęzyczna, angielsko-francuska.

Part 8: *Allocation of letters to the keys of a numeric keyboard*. Przyporządkowanie liter klawiszom klawiatury numerycznej (w telefonach i bankomatach); status tej części jest nieco inny niż poprzedniej, ale nie będziemy tutaj wchodzić w szczegóły.

Mówiąc w przybliżeniu, charakterystyczną cechą każdej współczesnej klawiatury alfanumerycznej jest przyporządkowanie każdemu klawiszowi dwóch znaków, przy czym wybór odpowiedniego znaku odbywa się za pomocą klawiszy nazywanych po angielsku *case shift*, a po polsku najczęściej *zmieniakiem rejestru*. Trudne do przetłumaczenia angielskie określenie wywodzi się podobno z czasów ręcznego składu tekstów, kiedy czcionki przechowywano w dwóch kasztach — majuskuły w kaszcie górnej (*upper case*), a miniskuły w dolnej (*lower case*). Norma ISO/IEC 9995 rozbudowuje ten mechanizm wprowadzając pojęcie poziomu (*level*). Dwa pierwsze poziomy odpowiadają dawnej dolnej i górnej kaszcie, zaś dawny klawisze *case shift* otrzymują nazwę *level 2 select*. Trzeci poziom jest włączany osobnym klawiszem, przy czym o ile DIS proponował konkretne położenie tego klawisza, wersja ostateczna podobno nic nie mówi na ten temat.

Innym istotnym pojęciem wprowadzonym przez normę ISO/IEC 9995 jest pojęcie *grupy* (ang. *group*). Sam termin nie jest najszczęśliwszy i podejrzewam, że brzmi on po angielsku równie niezręcznie, jak po polsku; sądzę, że przy ewentualnym tłumaczeniu normy na język polski nie należy tego terminu tłumaczyć dosłownie, lecz oddawać go np. przez *rozkład*. Zmiana grupy, dokonywana w bliżej nieokreślony sposób za pomocą odpowiedniego klawisza, powoduje zmianę funkcji całej klawiatury, w razie potrzeby łącznie z klawiszami funkcyjnymi. Liczba grup nie jest ograniczona. Norma nic nie stanowi na temat pierwszej, domyślnej grupy, która powinna być zgodna z lokalnymi normami lub zwyczajami. Zaleca się natomiast pewien specyficzny układ grupy drugiej klawiatury, który pozwala na wprowadzania znaków należących m.in. do 40 najważniejszych języków europejskich korzystających z alfabetu łacińskiego. Jest to niewątpliwie pożyteczna inicjatywa, która po jej upowszechnieniu pozwoli wielu użytkownikom wprowadzać teksty swojego języka w identyczny sposób niezależnie od miejsca pobytu.

Dla krajów wielojęzycznych bardzo istotna jest marginesowa dla nas część siódma, która pozwala ujednolić produkowane i stosowane klawiatury; np. w Kanadzie producenci powinni w zasadzie oferować aż 4 typy klawiatur — opisane tylko po angielsku, tylko po francusku, po angielsku z dodatkowym opisem francuskim, po francusku z dodatkowym opisem angielskim. Nawiasem mówiąc, liczba tego typu symboli zarejestrowanych przez ISO wynosi obecnie około 3000; jesteśmy więc na dobrej drodze do międzynarodowego języka ideograficznego.

4 Kody znaków

4.1 Kody 7-bitowe

Najpowszechniej stosowany kod 7-bitowy wywodzi się z ASCII — *American Standard Code for Information Interchange* — i tak też jest potocznie nazywany. Kod ten nabrał charakteru międzynarodowego po ustanowieniu normy ISO/IEC 646 *Information technology — ISO 7-bit coded character set for information interchange*, której ostatnie, trzecie wydanie jest datowane na 15.XII.1991. Polska nie miała dotąd bezpośredniego odpowiednika tej normy, ponieważ byliśmy zobowiązani do wdrażania norm EWG; tak więc polskie normy

PN-89/T-42108 Przetwarzanie informacji i komputery. Znaki alfanumeryczne. Klasyfikacja, nazwy i symbole.

PN-88/T-42109/01 Przetwarzanie informacji i komputery. Kod 7-bitowy. Tablica kodu i zestawy znaków SO i RWPG.

PN-88/T-42109/02 Przetwarzanie informacji i komputery. Kod 7-bitowy. Krajowe zestawy znaków.

PN-88/T-42109/02 Przetwarzanie informacji i komputery. Kod 7-bitowy. Krajowy zestaw znaków wprowadzany techniką rozszerzania kodu.

bazowały na odpowiednich normach RWPG i — jak się wydaje — nie miały większego wpływu na praktykę.

Polski odpowiednik normy ISO/IEC 646 przygotowano dopiero w 1992 roku i poddano go opiniowaniu przez 13 instytucji, wśród których był reprezentowany przeze mnie Instytut Informatyki UW. Nieliczne instytucje, które odpowiedziały na ankietę, zaopiniowały projekt pozytywnie; moim natomiast zdaniem tekst projektu zawierał nie tylko liczne niezręczności, ale i poważne błędy tłumaczenia zmieniające sens postanowień normy. Swoje stanowisko starałem się uzasadnić na konferencji uzgadniającej, która odbyła się 28.X.1992, w czasie której wstrzymałem się od głosowania nad nadaniem normie dalszego biegu. Mam wrażenie, że znaczna część moich uwag została uwzględniona. W szczególności zaakceptowano moją argumentację, że chociaż podstawowym znaczeniem angielskiego słowa *at* jest polskie *przy*, nazwa znaku `@` brzmiąca po angielsku *commercial at* powinna być tłumaczona jako *handlowe „po”*, a nie *handlowe „przy”*.

Dalsze losy tej normy nie są mi znane, być może zostanie ona wprowadzona w życie z dniem 1.I.1994.

4.2 Kody 8-bitowe

Problematyka kodów 8-bitowych jest bardzo obszerna, pominiemy więc tutaj wczesne normy, jak np.

PN-88/T-42112/01 Przetwarzanie informacji i komputery. Kod 8-bitowy. Tablica kodu i zestawy znakow ISO i RWPG.

które były oparte na normach RWPG, zajmiemy się natomiast dwoma głośnymi kiedyś kontrowersjami, które dobrze ilustrowały przysłowie *mądry Polak po szkodzie*.

W drugiej połowie lat osiemdziesiątych powstała wieloczęściowa norma ISO 8859 *Information processing. 8-bit single-byte coded graphic character sets*; każda z części opisuje jeden z kilku wzajemnie wykluczających się zestawów znaków. Zgodnie z obowiązującymi wówczas ustaleniami, w komitecie ISO opracowującym normę wszystkie kraje RWPG były reprezentowane przez delegację czechosłowacką. Prawdopodobnie za obopólnym porozumieniem przyjęto koncepcję, którą później nazwałem syndromem żelaznej kurtyny ([2]): część pierwsza normy definiuje kody języków zachodnioeuropejskich, część druga — języków krajów RWPG; ze względu na podział Niemiec znaki tego języka znalazły się w obu tabelach kodowych. Tak więc stosując normę ISO 8859-2 Czech może napisać list do Polaka, Słowaka lub Niemca, ale nie do Francuza czy Irlandczyka; analogiczne problemy występują przy stosowaniu pierwszej części normy czyli ISO 8859-1. Wszystkie dalsze problemy z polskimi literami w kodach komputerowych wydają się tylko konsekwencją ustaleń przyjętych w normie ISO 8859.

Z dniem 1.I.1991 weszła w życie polska norma PN-91/T-42115 będąca odpowiednikiem ISO 8859-2 w wersji z roku 1987. Choć decyzja ta w zasadzie była słuszna, jej forma miała prawo wzbudzić poważne zastrzeżenia. Norma ta została mianowicie wprowadzona jako obligatoryjna, określając zestaw znaków, który *należy stosować przy przetwarzaniu danych i obróbce tekstów oraz przy wymianie informacji*. Zgodnie z obowiązującym ustawodawstwem, użytkownicy komputerów IBM i kompatybilnych — zarówno osobistych, stosujących tzw. strony kodowe, o których będziemy mówić dalej, jak i „szafowych" (ang. *mainframe*) stosujących kod EBCDIC — zostali zagrożeni, jeśli dobrze rozumiem, karą aresztu do lat dwóch. Stworzenie takiej absurdalnej sytuacji na pewno nie podniosło autorytetu PKNMiJ ani w ogóle działalności normalizacyjnej w Polsce; uzytkownicy IBM będą mogli odetchnąć spokojnie dopiero 1.I.1995, kiedy wejdzie w życie ustawa z dnia 3 kwietnia 1993 r. o normalizacji (Dziennik Ustaw RP nr 55, poz. 251).

Norma PN-91/T-42115 spotkała się z powszechną krytyką w polskiej prasie informatycznej, ale często krytyka ta była bezprzedmiotowa, zarzucała bowiem, że norma ta nie stosuje się do komputerów osobistych typu PC. O ile mi wiadomo, kody stosowane na komputerach PC nigdy i nigdzie nie były zgłaszane jako normy krajowe czy międzynarodowe, pozostając w gruncie rzeczy prywatną sprawą producentów. Odpowiadając jednak na potrzeby użytkowników, wraz z systemem operacyjnym DOS 3.30 wprowadzono m.in. strony kodowe 850 i 852 bazujące odpowiednio na repertuarze kodów ISO 8859-1 i IS0 8859-2 — tak więc trudno mówić o narzucaniu strony kodowej 852, skoro norma ISO 8869-2 została ustanowiona przy kompletnym braku zainteresowania (co w tym przypadku oznacza akceptację) ze strony polskiego środowiska informatycznego.

O ile strona kodowa 850 była zawsze dostępna w standardowej dystrybucji systemu DOS, to w roku 1988 strona kodowa 852 (a także strona kodowa 855 z cyrylicą) była udostępniana użytkownikom PC-DOS (praktycznie sprzedawanego tylko z oryginalnymi komputerami IBM) w formie osobnej dyskietki *National*

Language Support. Od samego początku było możliwe stworzenie innej strony kodowej dla języka polskiego, ponieważ niezbędne informacje zostały opublikowane w *DOS Technical Reference manual* — o ile mi wiadomo, dla języka rosyjskiego pierwotnie stosowana strona kodowa została w praktyce wyparta przez tzw. alternatywny wariant Briabrina, podobny proces mógł mieć więc miejsce również dla języka polskiego. Jednak większość polskich użytkowników uświadomiła sobie istnienie strony kodowej 852 dopiero wtedy, kiedy pojawiła się ona w systemie operacyjnym i innym oprogramowaniu firmy Microsoft; stanowiano jej m.in. zarzuty wskazujące na to, że dyskutanci nigdy nie widzieli jej tabeli — np. twierdzono, że nie ma ona w ogóle znaków semigraficznych, podczas gdy pod tym względem jest ona identyczna ze stroną kodową 850, tzn. pominięto jedynie połączenia linii pojedynczych z podwójnymi. Starając się uniknąć tego typu nieporozumień opublikowałem pełną tabelę strony 852 w swoim artykule [1]; do dzisiaj jestem zdziwiony, że — o ile mi wiadomo — nikt tego nie zrobił przede mną.

Wśród różnych lokalnych sposobów kodowania polskich liter na PC największą popularność zdobył tzw. kod Mazovii. Nie będę tutaj powtarzał swojej opinii o nim, którą przedstawiłem we wspomnianym wcześniej artykule [1], chciałbym natomiast ustosunkować się do dalszych jego losów. Mianowicie z inicjatywy Andrzeja Gecowa otrzymał on od 1.IX.1992 status Normy Zakładowej ZN-92 *Przetwarzanie informacji. Zestaw znaków graficznych w jednobajtowym kodzie 8-bitowym — tzw. kod MAZOVIA* w Spółce Akcyjnej „Mikrokomputery". Wydaje mi się to wątpliwym osiągnięciem — takie rozwiązanie w niewielkim stopniu ułatwia dostęp do tekstu normy (ja swój egzemplarz otrzymałem bezpośrednio od Andrzeja Gecowa, za co mu przy okazji dziękuję), a status normy branżowej, której ważność wkrótce wygasa automatycznie, jest chyba słabym argumentem w negocjajach z zagranicznymi producentami. Warte rozważenia było chyba przyjęcie innej drogi, a mianowicie zarejestrowanie kodu Mazovii zgodnie z procedurą ustaloną w normie ISO 2375 *Procedure for Registration of Escape Sequences*. Tak czy inaczej, byłoby to jednak wyłamaniem się z powszechnego zwyczaju nienormalizowania w sposób formalny kodów PC, który to zwyczaj ma być może jakieś racjonalne uzasadnienie.

Ze względu na brak miejsca pominę tutaj aktualną kwestię kontrowersyjnej normy ISO/IEC 10367 *Information technology — Standardized coded graphic character sets for use in 8-bit codes* i jej polskiego odpowiednika PN- /T-42118 *Technika informatyczna. Znormalizowane zbiory znaków graficznych przeznaczone do stosowania w kodach 8-bitowych.*

Wszystkie wymienione wyżej kody są w zasadzie przeznaczone dla użytkowników korzystających przeważnie z jednego tylko języka. Krańcowo inne potrzeby występują np. przy wyszukiwaniu informacji bibliograficznych, gdzie użytkownik nie zawsze się orientuje, z jakiego języka pochodzą dane litery np. w nazwisku autora lub miejscu wydania. Do takich celów opracowano normę ISO 6937 *Coded Character Sets for Text Communication* składającą się z trzech części:

Part 1: General introduction.

Part 2: Latin Alphabetic and Non-Alphabetic Graphic Characters.
Part 3: Control Functions for Page-Image Format.

O praktycznym wykorzystaniu tego kodu w Polsce mówiono ostatnio w referacie [5].

Czytelników posiadających dostęp do poczty elektronicznej. a zainteresowanych problematyką wymienionych wyżej kodów, zachęcam do wzięcia udziału w elekronicznej liście dyskusyjnej *ASCII/EBCDIC character set related issues* `ISO8859@JHUVM.BITNET` lub do zapoznania się z archiwami tej listy.

4.3 Kod dwubajtowy (UNICODE)

Wielość różnorodnych kodów znaków oraz skomplikowane zależności między nimi, w szczególności możliwości dynamicznej zmiany kodu za pomocą tzw. sekwencji ucieczki (ang. *escape sequence*), stanowi poważne utrudnienie dla twórców oprogramowania o zasięgu światowym. Nic więc dziwnego, że kilka największych firm komputerowych, takich jak Apple, IBM, Microsoft i SUN, rozpoczęło prace nad rozwiązaniem tego problemu. W wyniku tych prac powstał projekt kodu dwubajtowego, w którym każdy znak jest reprezentowany przez 16 bitów. Aktualnie wykorzystuje się 28 000 pozycji do reprezentacji znaków różnych języków stosujących różne systemy pisma — łącznie z pismami ideograficznymi takimi jak japoński, chiński i koreański — zaś ponad 30 000 pozycji jest jeszcze niewykorzystanych. Pierwsza wersja UNICODE została opublikowana jako dwutomowe dzieło [7]. W związku z uznaniem UNICODE za podzbiór kodu wielobajtowego, omówionego w punkcie następnym, niezbędne były pewne zmiany, w wyniku których powstał kod UNICODE wersja 1.0.1. jako etap przejściowy do wersji 1.1. Wykaz zmian jest dostępny m.in. za pomocą sieci komputerowych pod adresem `Unicode.Org`.

4.4 Wielobajtowy kod uniwersalny

Podobne zadania, jak zespół UNICODE, postawiły sobie grupy robocze ISO, pracujące nad normą ISO/IEC 10646; w przeciwieństwie do UNICODE nie ograniczono reprezentacji znaku do dwóch bajtów, a opracowywany kod nazwano *uniwersalnym wielobajtowym kodowym zestawem znaków*. W trakcie pracy powstała bardzo ostra kontrowersja, czy kod ten ma zawierać w sobie UNICODE jako podzbiór. ciśle z tym związany był inny problem, mianowicie czy drobne warianty znaków ideograficznych mają być traktowane jako jeden znak czy nie. Ostatecznie zwyciężyła koncepcja UNICODE, który jako podzbiór uniwersalnego zestawu znaków otrzymał nazwę *podstawowej płaszczyzny wielojęzycznej*; 266 takich płaszczyzn może być w każdej tzw. grupie, których może być 64. W ogólnym wypadku znak w kodzie ISO 10646 reprezentowany jest przez ciąg czterech bajtów.

Ostateczny tekst tej normy został ustalony 11.III.1993. Jest to około 800stronicowy dokument ISO/IEC 10646-1 *Information technology — Universal Multiple-Octet Coded Character Set (UCS) — Part 1: Architecture and Basic Multilingual Plane*.

Osoby zainteresowane tą problematyką mogą zapisać się na elektroniczną listę dyskusyjną *Multi-byte Code Issues* `ISO10646@JHUVM.BITNET`, należy się liczyć jednak z faktem, że w czasie gorących kontrowersji ruch na liście jest bardzo duży.

5 Standaryzacja struktury tekstów (SGML)

Działalność normalizacyjna w zakresie języków naturalnych nie ogranicza się tylko do takich — pozornie prostych — zagadnień jak kody znaków czy klawiatury. Jednym z ciekawych zagadnień jest sformalizowanie opisu struktury tekstów. Problem ten wywodzi się z praktyki wydawniczej, kiedy przygotowując maszynopis do druku opatrywano go odpowiednimi adnotacjami dla składacza czy drukarni; po angielsku proces ten, jak i jego wynik, nosi nazwę *mark up* lub *markup*. W tradycyjnym procesie wydawniczym tego typu adnotacje miały charakter bardzo konkretny i operowały takimi pojęciami jak krój i stopień pisma drukarskiego, jednak już przy składzie komputerym takie adnotacje czy komendy mogą mieć charakter bardziej abstrakcyjny i zamiast np. operować stopniem pisma mogą tylko wskazywać, czy pismo ma być większe czy mniejsze od standardowego. Następnym krokiem abstrakcji jest wskazywanie tylko, z jakim elementem tekstu mamy do czynienia — tytułem, przypisem, tekstem właściwym itp. Tego typu adnotacje zasługują na nazwę uogólnionych (ang. *generalized*), stąd nazwa języka do reprezentacji struktury tekstu, a mianowicie *Standard Generalized Markup Language*. Został on wprowadzony normą ISO 8879 z 1986 r. oraz normami pochodnymi i pomimo silnego poparcia m.in. ze strony Departamentu Obrony Stanów Zjednoczonych nie zdobył sobie dotąd powszechnego uznania. Wydaje się jednak, że jego znaczenie systematycznie rośnie i zaczyna powoli osiągać masę krytyczną, która może spowodować gwałtowne przyspieszenie rozwoju i rozpowszechnienia oprogramowania wykorzystującego ten standard. Również dla tej problematyki istnieje elektroniczne forum dyskusyjne, jest ono jednak trudniej dostępne w Polsce ponieważ ma charakter tzw. *Usenet Newsgroup* o nazwie `comp.text.sgml`, zaś większość ośrodków w kraju nie ma jeszcze dostępu do tego typu informacji.

Jednym z najpoważniejszych i najbardziej znanych zastosowań SGML jest jego wykorzystanie jako podstawy tzw. *Text Encoding Initiative* czyli wspólnego przedsięwzięcia Stowarzyszenia dla Wykorzystania Komputerów w Humanistyce (*Association for Computers and the Humanities*), Stowarzyszenia Lingwistyki Obliczeniowej (*Association of Computational Linguistics*) i Stowarzyszenia Obliczeń Literackich i Lingwistycznych (*Association for Literary and Linguistic Computing*), zmierzającego do wypracowania uniwersalnych zasad kodowania tekstów i ich wymiany. Kolejny projekt tych zasad [6], oznaczony symbolem P2, jest dostępny m.in. za pomocą poczty elektronicznej pod adresem `LISTSERV@UICVM.BITNET`. Aktualne informacje i dyskusje o *Text Encoding Initiative* można znaleźć na elektronicznej liście dyskusyjnej `TEI-L@UICVM.BITNET`.

Przytoczony niżej przykład pochodzi z rozdziału 13 pracy [6], zatytułowanego *Base Tag Set for Terminological Data*.

```
<!-- Example 2a:  Nested Term Entry -->
<termEntry>
     <admin type='domain'> appearance of materials </admin>
     <tig lang=en>
          <term> opacity </term>
          <gram type=pos> n </gram>
          <descrip type='definition'> degree of obstruction to the
          transmission of visible light </descrip>
          <ptr type='bibliographic' target='ASTM.E284'>
          <admin type='responsibility' resp='E12'> </admin>
     </tig>
     <tig lang=de>
          <term> Opazit&auml;t </term>
          <gram type=pos> n </gram>
          <gram type=gen> f </gram>
          <descrip type='definition'> Ma&szlig; f&uuml;r die
          Lichtdurchsichtigkeit </descrip>
          <ref type='bibliographic' target='HFdn1983'> p. 383 </ref>
          <admin type='responsibility' resp='DIN TC for paper
          products'></admin>
     </tig>
     <tig lang=fr>
          <term> opacit&eacute; </term>
          <gram type=pos> n </gram>
          <gram type=gen> f </gram>
          <descrip type='definition'> rapport du flux lumineux
          incident au flux lumineux transmis ou r&eacute;fl&eacute;chi
          par un noircissement photographique </descrip>
          <ptr type='bibliographic' target='HJdi1986'>
          <admin type='responsibility' resp='C.I.R.A.D.'> </admin>
     </tig>
</termEntry>
```

Jak łatwo zauważyć, informacja w powyższym przykładzie jest zorganizowana hierarchicznie za pomocą *oznaczników* wskazujących początek i koniec *elementu* danego typu; termin *oznacznik* jako tłumaczenie angielskiego terminu *tag* przyjmuję za słownikiem [4]. Wśród elementów można zauważyć *elementy wiążące* (ang. *linking elements*); jeśli odsyłamy do kompletnego opisu, stosujemy element `ptr` (ang. *pointer*, odsyłacz), jeśli zaś do jego fragmentu — element `ref` (ang. *reference*, przywołanie); w obu wypadkach cytowany opis bibliograficzny jest oznaczony symbolem utworzonym zgodnie z dokumentem ISO/TC 37 WI 18 *Coding of Bibliographic References in Terminology Work and Terminography* (1991). Cały zapis wykorzystuje tylko znaki ASCII (ISO 646), ale dowolny znak można zapisać stosując konstrukcję *przywołania całostki* (ang. *entity reference*); w naszym przykładzie litera *é* (*e* z akutem) jest zapisywana jako `&eacute;`, litery *ä* i *ü* (*a* i *u* z przegłosem) jako `&auml;` i `&uuml;`, zaś litera *ß* („długie *s*") jako `&szlig;`.

Poszczególne elementy mogą posiadać różne atrybuty. Element `tig` (ang. *term information group*) zawiera atrybut `lang` (ang. *language*), którego wartość

wskazuje na język danego terminu za pomocą symbolu języka zdefiniowanego w normie ISO 639; element `gram` zawiera atrybut `pos` (ang. *part of speech*) itd.; zestaw elementów i atrybutów dla dokumentów terminologicznych jest opracowywany w ISO przez grupę roboczą WG 1 podkomitetu SC 3 komitetu technicznego TC 37 i przewidziany do opublikowania jako norma ISO 12620.

W powyższym przykładzie SGML określa jedynie formalną składnię tego zapisu. W niektórych przypadkach odbywa się to bezpośrednio, np. postać przywołania całostki — znak ampersand, nazwa całostki i średnik — jest określona przez SGML. Bardziej wyrafinowane zależności składniowe są jednak określone tylko pośrednio, mianowicie SGML definiuje składnię i znaczenie *opisu typu dokumentu* (ang. *Document Type Description*, DTD), a podstawowym zadaniem TEI jest właśnie sformułowanie składni różnego typu dokumentów za pomocą odpowiednich DTD i przypisanie poszczególnym jednostkom składniowym odpowiedniego znaczenia.

Powyższy przykład ilustruje również inną cechę SGML i TEI — w swojej oryginalnej postaci są one przeznaczone bardziej dla komputerów niż dla ludzi ze względu na rozwlekły i redundantny zapis oznaczników. Chociaż SGML przewiduje pewne metody uproszczania tych zapisów, moim zdaniem szerokie rozpowszechnienie się tych standardów wymaga uprzedniego pojawienia się narzędzi programistycznych pozwalających manipulować zapisami SGML w sposób bardziej wygodny dla użytkownika.

6 Uwagi końcowe

Swój apel o większe zainteresowanie problematyką normalizacji chciałbym wzmocnić konstruktywnym akcentem podając adres (na dzień 15.XI.1993) Sekretariatu Normalizacyjnej Komisji Problemowej do spraw Informatyki: *Instytut Maszyn Matematycznych, ul. Krzywickiego 34, 02-078 Warszawa*. Posiedzenia — niekiedy bardzo długie — odbywają się z reguły raz w miesiącu; członkom (nominowanym na wniosek NKPI przez Prezesa PKNMiJ) oraz zaproszonym gościom przysługuje za każde posiedzenie ryczałtowe wymagrodzenie w wysokości około 30 tysięcy zł, które w praktyce i tak nie jest wypłacane; wiąże się to z założeniem, że prace normalizacyjne powinne być finansowane przez zainteresowane instytucje i przedsiębiorstwa, które delegują swoich pracowników do NKPI na swój koszt — mam poważne wątpliwości, czy założenie to sprawdzi się na dłuższą metę. W związku z wchodzeniem w życie nowej ustawy o normalizacji czeka nas zapewne więcej różnych eksperymentów. Jedną z intrygujących dla mnie nowości są np. Z*asady opracowywania i sposób prezentacji polskich norm stanowiących wprowadzenie norm europejskich*, które stanowią w punkcie 2.2 (jeśli dobrze rozumiem), że spis treści nie może zawierać numerów stron; niepokojące jest tutaj nie tylko to, że korzystanie np. z polskiego tłumaczenia 800stronicowej normy ISO 10646 bez porządnego spisu treści wydaje się raczej niewygodne, ale również to, że tego typu zasady nie są w ogóle dyskutowane na NKPI ani tym bardziej na forum publicznym. Inną nowością, stanowiącą moim zdaniem krok wstecz, jest zlikwidowanie tzw. norm okładkowych; były to polskie

odpowiedniki norm zagranicznych ustanawiane bez tłumaczenia oryginalnego tekstu, lecz opatrzone tylko okładką z polskim tytułem i innymi informacjami. Normy te praktycznie nie były stosowane z bardzo prozaicznego względu — PKNMiJ finansował branżowe ośrodki normalizacyjne proporcjonalnie do objętości przygotowywanych lub tłumaczonych norm, a więc na normach okładkowych nie można było zarobić; normy takie są jednak cały czas w użyciu w kilku krajach zachodnich i w wielu wypadkach są optymalnym rozwiązaniem — tłumaczenie na język narodowy kilkusetstronicowych norm informatycznych po to, aby je przeczytało kilku specjalistów, którzy i tak znają język oryginału, jest ewidentnym marnowaniem czasu i środków. Reasumując, *jak sobie pościelisz, tak się wyśpisz* i jeśli kompetetentne środowiska będą systematycznie ignorować problematykę normalizacyjną, to trudno spodziewać się wysokiego poziomu ustanawianych norm — w najlepszym razie reprezentować one będą nie interes całego społeczeństwa, lecz tylko tych instytucji, które stać na luksus oddelegowania swoich pracowników do NKPI, jej grup roboczych i międzynarodowych organizacji normalizacyjnych.

Literatura

[1] Janusz S. Bień. Polskie litery na PC (głos w dyskusji). *ComputerWorld PL* nr 4(10), 16.II.1991, s. 9-11,14,16,19,21.

[2] Janusz S. Bień. Strona kodowa 852 i syndrom żelaznej kurtyny. *Biuletyn Polskiego Towarzystwa Informatycznego* r. X nr 5, s. 3, 1991.

[3] Piotr Carlson, Marek Urbański. Minimaliści i maksymaliści. *PCkurier* 1/93 (7 stycznia 1993), s. 63–65.

[4] Andrzej Marciniak, Michał Jankowski. **Słownik informatyczny angielskopolski**. Państwowe Wydawnictwo Naukowe, Warszawa–Poznań 1991.

[5] Janusz J. Młodzianowski. National Character Support in Telnet. Network Services Conference '93, Warsaw, Poland, 12–14 October 1992, Booklet of abstracts p. 54.

[6] C. M. Sperberg–McQueen, Lou Bernard (eds.). **Guidelines for Electronic Text Encoding and Interchange**. Draft Version 2. Text Encoding Initiative, Chicago, Oxford, 1993.

[7] The Unicode Consortium. **The Unicode Standard: Worlwide Character Encoding, Version 1.0**. Addison-Wesley 1991.