INSTYTUT SLAWISTYKI POLSKIEJ AKADEMII NAUK
FUNDACJA SLAWISTYCZNA

SEMANTYKA A KONFRONTACJA JĘZYKOWA

tom 5

Pod redakcją
Danuty Roszko i Joanny Satoły-Staśkowiak

Warszawa 2015

Tom opiniowały do druku
Anna Krupska-Perek, Dorota Urbanek

**Wydanie publikacji dofinansowane
przez Ministerstwo Nauki i Szkolnictwa Wyższego**

Redakcja
Ewa Dzierżanowska, Dorota Rdest

Skład
DTP Beata Jankowska

ISBN 978-83-64031-25-0

Slawistyczny Ośrodek Wydawniczy
Instytutu Slawistyki PAN
ul. Jaracza 6 m. 12, 00-378 Warszawa
sow@ispan.waw.pl www.ispan.waw.pl

SPIS TREŚCI

JANUSZ S. BIEŃ
Katedra Lingwistyki Formalnej
Wydziału Neofilologii Uniwersytetu Warszawskiego
Warszawa

PROBLEMY KODOWANIA ZNAKÓW W KORPUSACH HISTORYCZNYCH

1. Wstęp

Tytułowy problem sygnalizował już Krzysztof Opaliński w artykule [Opaliński 2007], tutaj przedstawimy go jednak bardziej szczegółowo na podstawie doświadczenia zdobytych przy digitalizacji słowników historycznych [por. Bień 2012c] i przy pracy nad udostępnieniem tzw. korpusu IMPACT [por. Bień 2014] – praktycznie pierwszego[1] polskiego korpusu tekstów dawnych – i związanych z tym przemyśleń.

2. Pojęcie znaku w standardzie Unicode

Przypomnę tutaj kilka faktów, o których pisałem już przy innych okazjach. Współcześnie teksty w komputerach zapisywane są niemal wyłącznie w standardzie Unicode (http://www.unicode.org/) – aktualna wersja to 8.0.0 z lipca 2015 r. Podstawową jednostką tekstu w tym standardzie jest *znak kodowy* (ang. *coded character*) – obiekt abstrakcyjny, w praktyce trzeba traktować go jako zdefiniowany przez wyliczenie (ich liczba przekracza 100 tysięcy). Znaki są identyfikowane przede wszystkim przez przyporządkowane im jednoznacznie liczby nazywane *współrzędnymi kodowymi* (ang. *code points*); zgodnie z obowiązującą konwencją liczby te są zapisywane w systemie szesnastkowym, czasami – dla wskazania, że chodzi o współrzędną kodową – poprzedzane prefiksem U+, np. U+0020. Dodatkowo

[1] *Korpus tekstów staropolskich do roku 1500*, stanowiący wynik realizowanego w latach 2000–2003 w Instytucie Języka Polskiego PAN grantu KBN 1 H01D 018 19, jest bardzo mały i pozbawiony jakichkolwiek funkcji wyszukiwania, por. https://www.ijp-pan.krakow.pl/publikacje-elektroniczne/korpus-tekstow-staropolskich

znaki są jednoznacznie identyfikowane przez umowne – często nieprecyzyjne, a czasami nawet mylące – nazwy w języku angielskim, np. DIAMETER SIGN (U+2300) lub LATIN CAPITAL LETTER O WITH STROKE (U+00D8).

Znaki mają *semantykę* (termin rozumiany tutaj bardzo wąsko i w konsekwencji mylący) określoną przez zestaw formalnych *własności* o ściśle określonych nazwach i wartościach. Własności te dotyczą jednak prawie wyłącznie kwestii czysto technicznych. Wygląd znaku i jego przeznaczenie są w standardzie określone tylko pośrednio i nieprecyzyjnie przez podanie jego przykładowego kształtu i ewentualnie nieformalne komentarze.

Kształt znaku jest technicznie nazywany *glifem* (ang. *glyph*). Pierwsze jego użycie w interesującym nas znaczeniu pojawiło się chyba w międzynarodowej normie ISO/IEC 9541-1:1991 *Information technology – Font information interchange*. W moim swobodnym tłumaczeniu podana tam definicja brzmi: *Abstrakcyjny symbol graficzny, którego kształt jest określony w stopniu pozwalajacym na jego rozpoznanie i identyfikację, ale bez przesądzania konkretnych cech jego wyglądu*.

W praktyce terminu tego używa się często na oznaczenie znaków piśmiennych dostępnych w konkretnym *foncie* (ang. *font*), czyli mniej lub bardziej konkretnym kroju pisma drukarskiego – w ogólnym przypadku znaki te rzeczywiście mogą stanowić abstrakcyjne wzorce ulegające różnym modyfikacjom (np. powiększeniu lub pochyleniu) już na etapie drukowania czy wyświetlania. Nie był jakoś dotąd odczuwany brak terminu oznaczającego konkretny kształt znaku – proponujemy zaadaptować do tego celu termin *czcionka* (ang. *type* lub *sort* – oba terminy niestety niewygodne ze względu na ich wieloznaczność, zwłaszcza w kontekstach informatycznych).

W wielu przypadkach nie da się obiektywnie stwierdzić, czy dwóm różnym konkretnym czcionkom odpowiadają dwa glify, czy dwa znaki kodowe – ostatecznie musi decydować mniej lub bardziej arbitralna konwencja.

W standardzie Unicode pewna pula współrzędnych kodowych jest wydzielona jako tzw. obszar użytku prywatnego (ang. *Private Use Area*, w skrócie PUA), który może być wykorzystany do kodowania dowolnych znaków na zasadzie porozumienia zainteresowanych stron. Dla nas szczególnie istotna jest propozycja kodowania znaków niedostępnych we właściwym standardzie sformułowana w rekomendacjach *Medieval Unicode Font Initiative* – w skrócie MUFI. Aktualna wersja rekomendacji nosi numer 3.0 i została opublikowana w 2009 r. (por. http://www.mufi.info/). Współrzędne kodowe znaków rekomendowanych przez MUFI będziemy poprzedzać prefiksem M+.

Standard Unicode umożliwia tworzenie znaków złożonych (ang. *composite characters*, co proponuję tłumaczyć jako *znaki kompozytowe* lub krótko *kompozyty*) przez dodawanie diakrytów – ogólniej *znaków dostawnych* (ang. *combining charakters*) – do wybranego znaku stanowiącego wówczas *znak bazowy* (ang. *base*

character). Konsekwencją tego jest fakt, że niektóre znaki mogą być reprezentowane na kilka sposobów, np. litera *ń* może być zapisana jako kompozyt (U+0144) albo jako sekwencja dwóch znaków: litery *n* (U+006E) i akcentu akutowego (U+0301) – sekwencję taką zapisujemy formalnie <U+006E,U+0301>.

Niektóre znaki to ligatury, np. LATIN SMALL LIGATURE FI (ﬁ, U+FB01) w Unikodzie i LATIN SMALL LIGATURE LONG S DESCENDING T (ﬅ, M+EADA) w rekomendacji MUFI; na potrzeby wyszukiwania często właściwe jest rozłożenie ligatur na elementy składowe.

Reasumując, pojęcie znaku w Unikodzie nie jest zbyt wygodne z praktycznego punktu widzenia, w dodatku – zwłaszcza w języku polskim – termin *znak* jest kłopotliwy ze względu na inne znaczenia. W związku z tym proponuję obiekty reprezentowane w standardzie Unicode przez współrzędne kodowe nazywać *tekstonami* (to sformułowanie uwzględnia pewne subtelności standardu, które tutaj nie są omawiane). Swoją drogą już od pewnego czasu proponuję termin *tekstel* na oznaczenie podstawowych jednostek tekstu bez względu na ich reprezentację w standardzie Unicode.

Warto wspomnieć, że standard Unicode posługuje się również pojęciem *user-perceived character*, w swobodnym tłumaczeniu *znak w rozumieniu użytkownika*. Znaki takie są reprezentowane przez sekwencje znaków Unicodu tworzące *zbitki grafemiczne* (ang. *grapheme cluster*), termin grafem (*grapheme*) nie jest używany celowo. Zbitki grafemiczne nie są jednak definiowane przez standard, który formułuje tylko pewne ogólne reguły i podaje kilka przykładów, głównie z języków orientalnych. Jak się wydaje, utożsamianie teksteli ze zbitkami grafemicznymi nie jest właściwe, ale sprawa wymaga jeszcze szczegółowej analizy.

3. Repertuary czcionek, tekstonów, teksteli i glifów

3.1. Czcionki

W roku 1920 Ludwik Bernacki (dyrektor Zakładu Narodowego im. Ossolińskich we Lwowie) zaproponował plan wydawniczy, którego jednym z elementów miał być *atlas zupełnego zasobu typograficznego oficyn ówczesnych* – tj. XVI i XVII wieku – *na ziemiach Rzeczypospolitej* [por. Piekarski 1936]. Postulat ten próbował zrealizować w 1936 r. Kazimierz Piekarski (bibliotekarz, bibliograf, historyk książki), publikując – częściowo własnym kosztem – „Polonia typographica saeculi sedecimi". Po II wojnie światowej jego współpracowniczka Alodia Kawecka-Gryczowa wydała kilka kolejnych zeszytów serii, ale po jej śmierci nikt nie podjął się kontynuacji. Publikacje te są trudno dostępne ze względu na niski nakład; Katedra Lingwistyki Formalnej UW dysponuje na własne potrzeby skanami praktycznie wszystkich części, ale ze względu na niejasne prawa autorskie nie może ich udostępnić publicznie. Wyjątek stanowi zeszyt pierwszy, do którego prawa wygasły w roku

2014 (ponieważ minęło 70 lat od śmierci autora) i który w związku z tym mógł zostać udostępniony (http://teksty.klf.uw.edu.pl/40, adres może ulec zmianie).

Innym, bardziej pierwotnym, źródłem informacji o dawnych czcionkach są oryginalne wzorniki czcionek poszczególnych drukarń – warto je chyba zebrać w publicznie dostępną kolekcję. Katedra może wnieść niewielki wkład do takiego projektu, ponieważ dysponuje skanami dwóch wzorników ze zbiorów BUW (http://teksty.klf.uw.edu.pl/1/ i http://teksty.klf.uw.edu.pl/2/, adresy mogą ulec zmianie).

W sytuacji, kiedy brakuje syntetycznego opisu repertuaru czcionek, interesujące są narzędzia pozwalające uzyskać tę informację dla konkretnej publikacji. Próba stworzenia takich narzędzi została podjęta w ramach grantu *Narzędzia digitalizacji tekstów na potrzeby badań filologicznych* (por. sprawozdanie końcowe [Bień 2012b]), podobne próby są również podejmowane w ramach projektów zagranicznych, por. np. projekty PaRADIIT (https://sites.google.com/site/paradiitproject/) i IMPACT (http://www.digitisation.eu/tools/browse/experimental-prototypes/inventory-extraction). Specyficzną cechą naszego podejścia – niestety niezrealizowanego w całości – było zastosowanie architektury klient–serwer, co pozwala na tworzenie wspólnej, sukcesywnie uzupełnianej bazy czcionek dostępnej w Internecie.

3.2. Tekstony

Standard Unicode definiuje obiekty, które nazwaliśmy tutaj tekstonami, na dwa sposoby. Pierwszy i w gruncie rzeczy podstawowy sposób to zestaw plików komputerowych określających nazwy i własności tych obiektów. Jest on nazywany *Unicode Character Database*, czyli baza danych znaków Unicodu, choć nie jest to baza danych w normalnym znaczeniu tego słowa, lecz pliki tekstowe (ang. *plain text*).

Jeden z tych plików służy do tworzenia czytelnych dla człowieka tabel w postaci dokumentu w bardzo popularnym formacie PDF. Dokument ten jest tworzony za pomocą bezpłatnego programu Unibook Character Browser (http://www.unicode.org/unibook/), ale do jego stworzenia niezbędne są fonty zawierające odpowiednie glify, które już bezpłatne nie są. W praktyce samodzielne stworzenie takiego dokumentu okazuje się nieopłacalne, zwłaszcza że jest on dostępny bezpłatnie w Internecie.

Potrzeba wyszukiwania tekstonów na podstawie ich nazwy lub własności jest powszechna w informatyce i istnieje wiele narzędzi służących do tego celu. Są to zarówno proste programy instalowane na komputerze użytkownika, jak i mniej lub bardziej wyrafinowane wyszukiwarki w postaci witryn WWW; osobiście często korzystam z witryny http://www.fileformat.info/info/unicode/.

Repertuar tekstonów MUFI został pierwotnie opisany w przygotowanych ręcznie eleganckich dokumentach w formacie PDF, które jednak okazały się trudne do uaktualniania. W ramach projektu ENRICH (*European Networking Resources and Information concerning Cultural Heritage* stworzono witrynę *Gaiji Bank: data-*

base of non-standard characters (http://www.manuscriptorium.com/apps/gbank/gbank_table.php) prezentującą tekstony MUFI w nieco innej formie. Nie jest w tej chwili przesądzone, jaką formę będzie miała kolejna specyfikacja MUFI. Byłoby bardzo wskazane, gdyby pod względem formalnym przypominała ona bardziej standard Unicode.

W nowej specyfikacji MUFI powinien znaleźć się zgłoszony przeze mnie tekston LATIN SMALL LIGATURE LONG S L WITH STROKE, por. http://www.mufi.info/pipeline/for-v4.html i http://www.mufi.info/pipeline/.

Warto dodać, że *gaiji* to japoński termin przejęty przez konsorcjum *Text Encoding Initiative* na oznaczenie znaków spoza standardu Unicode. Rekomendacje konsorcjum TEI przewidują możliwość opisu nietypowych znaków właśnie za pomocą elementu *gaiji*, nie jest to jednak satysfakcjonujące [por. dyskusję w artykule Fredell, Borchers IV, Ilgen 2013] .

3.3. Tekstele

Do opisu teksteli można wykorzystać pewne mechanizmy przewidziane dla zbitek grafemicznych, mimo że te pojęcia uważamy za różne (choć oczywiście podobne). Tym mechanizmem jest wykaz tzw. *named sequences*, czyli sekwencji znaków posiadających własne nazwy. W aktualnej wersji standardu wykaz ten liczy tylko kilkadziesiąt pozycji dobranych według niejasnych kryteriów: znajdujemy tam zarówno LATIN SMALL LETTER O WITH VERTICAL LINE BELOW AND GRAVE, jak i TAMIL SYLLABLE KSSA. Najważniejsza dla nas jednak jest nie jego treść, ale forma – w analogiczny sposób możemy definiować interesujące nas tekstele.

Standard zawiera również niewielki wykaz aliasów, czyli alternatywnych nazw znaków. Z technicznego punktu widzenia nie jest on niezbędny, jeśli dopuścimy *named sequences* jednoelementowe, co na obecnym etapie wydaje się wygodniejsze.

W 2009 r. jeden z moich studentów na moją prośbę przygotował rozszerzoną wersję znanego programu unihist (http://billposer.org/Software/unidesc.html) do tworzenia histogramów znaków w tekście zakodowanym zgodnie ze standardem Unicode. Najważniejszym rozszerzeniem jest uwzględnianie w statystyce znaków złożonych i nadawanie im nazw za pomocą pliku w formacie NamedSequences.txt. Przykładem wyniku działania programu jest histogram wykonany dla korpusu IMPACT, dostępny jako załącznik do notatki [Bień 2012a].

Program ten jest dostępny na zasadach swobodnej licencji GNU, aktualnie pod adresem https://bitbucket.org/jsbien/unihistext.

3.4. Glify

Podstawową formą repertuaru glifów są fonty. Choć historycznie jednemu tekstonowi w konkretnym odpowiadał jeden glif, obecnie sytuacja jest bardziej

skomplikowana – sekwencja kilku tekstonów może być reprezentowana przez jeden glif, a także jeden tekston może być alternatywnie reprezentowany przez kilka różnych glifów [Haralambous 2007]. Stwarza to problemy decyzyjne, czy różne czcionki traktować jako różne tekstony, czy różne glify. Przykład stanowią wahania, jak traktować ligatury występujące w *Nowym Karakterze Polskim* [Januszowski 1594] – por. http://www.mufi.info/pipeline/ punkt 13.

Istnieje kilka fontów uwzględniających w części lub w całości rekomendacje MUFI o różnym statusie prawnym. Na szczególną uwagę zasługuje font Junicode (*Julius Unicode*) udostępniony na swobodnej licencji pozwalającej na własne modyfikacje. Istnieje wiele programów prezentujących na różne sposoby zawarte w foncie glify, przeznaczonych zarówno dla zwykłych użytkowników, jak i dla twórców fontów.

Jednym z takich programów jest fntsample (http://fntsample.sourceforge.net/), tworzący tabele glifów na wzór standardu Unicode. Jeden z moich studentów na moją prośbę rozszerzył ten program o pewne funkcje wspomnianego wcześniej programu Unibook, a mianowicie o możliwość dodawania komentarzy. Przykład wyniku działania tego programu znajduje się pod adresem https://bitbucket.org/jsbien/parkosz-font/downloads/Parkosz1907draft.pdf. Sam program jest dostępny na swobodnej licencji i może być nadal rozwijany, aktualnie znajduje się m.in. pod adresem https://github.com/ppablo28/fntsample_ucd_comments.

4. Optyczne rozpoznawanie znaków

Do niedawna jedynym sposobem wprowadzenia dawnego tekstu do klawiatury było ręczne przepisanie go. Obecnie coraz częściej próbuje się wykorzystać do tego celu programy optycznego rozpoznawania znaków – przykład może stanowić projekt *Early Modern OCR* (http://emop.tamu.edu/). Programy takie poddaje się „trenowaniu" na odpowiednich tekstach wzorcowych. Obszerny zbiór takich tekstów wzorcowych został przygotowany w ramach projektu IMPACT i udostępniony m.in. w formie wspomnianego już korpusu. W tekstach tych konieczne było reprezentowanie znaków w sposób maksymalnie zgodny z ich wyglądem, stąd np. znak diakrytyczny nad literą *z* był kodowany – zgodnie z jego faktycznym wyglądem – również jako tylda (*z̃*), „haczek" (ang. *caron*, *ž*) czy makron (*z̄*).

W ramach tego projektu przygotowano również eksperyment porównujący komercyjny program FineReader i bezpłatny swobodny program Tesseract [Heliński, Kmieciak, Parkoła 2012]. Zostały udostępnione publiczne dane wykorzystywane do testów i wynik trenowania programu Tesseract, który może być wykorzystywany do rozpoznawania innych podobnych tekstów – por. http://dl.psnc.pl/activities/projekty/impact/results/. Nie wiem jednak, w jakim stopniu dane te są wykorzystywane w praktyce i czy uzyskane wyniki są zadowalające.

5. Transkrypcje i transliteracje

Pojęcie transkrypcji rozumiemy szeroko, transliterację traktujemy jako szczególny przypadek transkrypcji. Za pracą [Driscoll 2006] będziemy mówić o różnych poziomach transkrypcji. Tradycyjnie najbardziej wierna transkrypcja nosi nazwę transkrypcji dyplomatycznej (ang. *diplomatic transcription*) – termin pochodzi od dyplomatyki, nauki pomocniczej historii zajmującej się dokumentami. Na potrzeby trenowania programów do rozpoznawania znaków potrzebny jednak jest jeszcze bardziej wierny zapis tekstu – nazywam go transkrypcją faksymilową.

W przypadku korpusów dobór transkrypcji nie jest tak istotny, jak w przypadku przygotowywania tekstu do druku, ponieważ w pewnych granicach zmiany transkrypcji mogą być dokonywane automatycznie na etapie tworzenia korpusu lub na etapie formułowania kwerend przez użytkownika. W konsekwencji na etapie wprowadzania tekstów lepiej stosować transkrypcję wierną, ewentualna modernizacja pisowni może być dokonana później.

W ostatnim czasie pojawiły się dwa programy do tego celu. Jeden został zrealizowany pod moim kierunkiem na potrzeby projektu IMPACT; razem z przykładowymi regułami jest on dostępny na swobodnej licencji GNU pod adresem https://bitbucket.org/jsbien/pol. Drugi został zrealizowany na potrzeby projektu SYNAT[2] i jest swobodnie dostępny pod adresem https://github.com/kawu/hist-pl. Porównanie obu programów to jeszcze kwestia przyszłości.

6. Teksty i metateksty

Zasadniczy problem stanowią teksty z okresu, kiedy nie ukształtowały się zasady pisowni polskiej. Ważnym tekstem tego typu jest rękopiśmienny łaciński traktat Parkosza o ortografii polskiej. Choć jego propozycje nigdy nie weszły w życie, trzeba umieć o nich pisać, a w konsekwencji trzeba dysponować jakąś graficzną reprezentacją proponowanych przez niego znaków. Jeden z moich studentów na moją prośbę przygotował specjalny font, dostępny pod adresem https://bitbucket.org/jsbien/parkosz-font-old i zawierający takie wymyślone przez Parkosza glify, jak  i  .

Otwarta pozostaje jednak sprawa przyporządkowania tym znakom powszechnie akceptowanych współrzędnych kodowych. Moja aktualna propozycja zawarta jest niejawnie w tworzonym przeze mnie foncie, dostępnym pod adresem https://bitbucker.org/jsbien/parkosz-font.

Na uwagę zasługuje również *Raj duszny*, do niedawna uważany za pierwszą polską książkę. Do dzisiaj żaden egzemplarz się nie zachował, ale znanych jest kilka stron dzięki reprodukcjom w książce [Bernacki 1918]. Samogłoski nosowe są

[2]SYstem NAuk i Techniki, projekt Programy Strategiczne NCRiR, pt. „Utworzenie uniwersalnej, otwartej, repozytoryjnej platformy hostingowej i komunikacyjnej dla sieciowych zasobów wiedzy nauki, edukacji i otwartego społeczeństwa wiedzy" o budżecie około 60 mln zł, realizowany w latach 2010–2013.

tam oznaczane przez literę α – samodzielną lub z kropką pod spodem. Konwencję tę zachowuje w cytatach „Słownik polszczyzny XVI wieku". Przykłady innych nieużywanych współcześnie liter można znaleźć m.in. w książce [Górski et al. 1955].

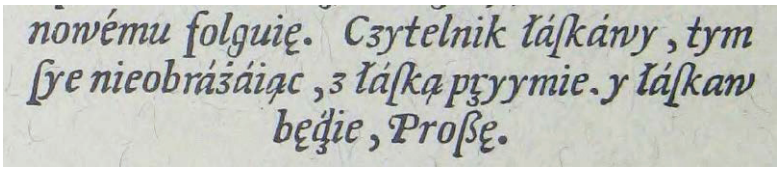
Jak wiadomo, dawna polszczyzna rozróżniała samogłoski jasne i pochyłe, ale rozróżnienie to nie było zaznaczane w druku konsekwentnie, stąd np. „Słownik polszczyzny XVI wieku" stosuje *å* na oznaczenie *a* pochyłego niezależnie od tego, jak jest reprezentowane w piśmie (jak się wydaje, konwencja ta ma dłuższą tradycję, ale nie udało mi się ustalić jej pochodzenia). Znak ten nie stwarza jednak problemu, gdyż jest dostępny w Unikodzie i wielu fontach.

Bardziej kłopotliwa jest kwestia „metaznaków" wprowadzonych w „Słowniku polszczyzny XVI wieku". W instrukcji redakcyjnej w ustępie pod tytułem „Znaki dla zapisów fonetycznie dwuznacznych" (http://www.spxvi.edu.pl/instrukcja/V/39/) czytamy: *W wypadku rozpatrywania pewnej cechy fonetycznej, zwłaszcza wchodzącej w oboczności, stosujemy dla użyć pochodzących z tekstów, które nie dostarczają danych co do brzmienia omawianego zjawiska, znak litery w nawiasie.*

Możemy więc spotkać następujące symbole reprezentujące pojedyncze litery: *(a)*, *(e)*, *(o)*, *(ę)*, *(ø)*, *(s)*, *(n)*, *(b)*, *(p)*, *(w)*, *(m)* i nawet *(cz)*. W dodatku: *Drugą funkcją nawiasu w części gramatycznej jest oznaczanie fakultatywnych cząstek wyrazu, tj. takich, które będąc dla rozpatrywanego właśnie zagadnienia zjawiskiem drugorzędnym, mogą w danej formie występować lub nie […].*

Rozróżnienie obu użyć nie zawsze jest łatwe. Inspiracją dla proponowanego przeze mnie rozwiązania tego problemu jest Unikodowy zestaw „znaków otoczonych", zawierających m.in. małe litery łacińskie w nawiasach okrągłych (stosowane w nieznanej mi funkcji w tekstach orientalnych), np. ⒜, czyli PARENTHESIZED LATIN SMALL LETTER A (U+249C). Wydaje się, że zestaw ten warto uzupełnić tak, aby wszystkie znaki fonetycznie dwuznaczne były pojedynczymi znakami w sensie Unikodu.

Tekstem zasługującym na szczególną uwagę jest XVI-wieczny traktat ortograficzny „Nowy Karakter Polski" [Januszowski 1594] porównujący kilka wariantów pisowni polskiej. Tekst należy do kanonu źródeł „Słownika polszczyzny XVI wieku" (czyli podlegał ekscerpcji pełnej) i jest w słowniku cytowany, jednak – jak sygnalizowałem to w artykule [Bień 2011] – niestety niekonsekwentnie, częściowo z powodu ograniczeń technicznych. W wyszukiwarce leksykograficznej Katedry Lingwistyki Formalnej (korpusy.klf.uw.edu.pl) słownik dostępny jest w takiej formie, w jakiej został przekazany do drukarni – można zauważyć np. że ligatura *dź* specyficzna dla tego utworu została zbudowana z wykorzystaniem pomniejszonej litery *з* (CYRILLIC SMALL LETTER ZE, U+0437), por. ilustracje 1. i 2. W rezultacie zawierające ją słowo jest praktycznie nie do odnalezienia (w wyszukiwarce reprezentują je trzy elementy: bę'd, *з* i ie).

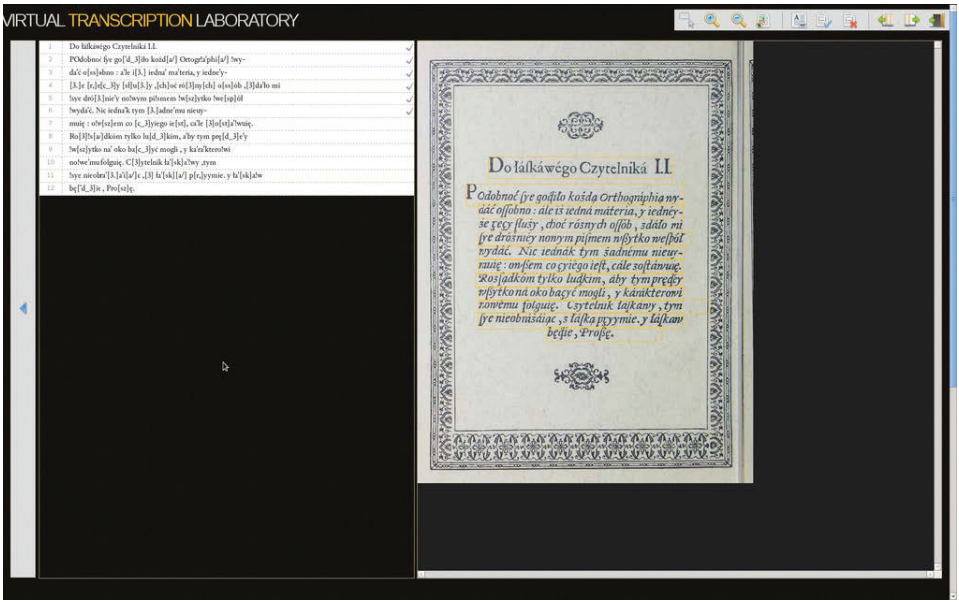


Rysunek 1. „Nowy Karakter Polski", arkusz A, karta 1, verso

KochFrag 4; *KołakCathOkuń* A2v; Czytelnik łáſkáwy [...] z łáſką pzyymie, y łáſkaw będzie *JanNKar* Av, A4v; *SarnStat* 330; *Sieb-*

Rysunek 2. „Słownik polszczyzny XVI wieku", tom XXXIII, s. 273

W celu dokładniejszej analizy typograficznej tego tekstu został on umieszczony przeze mnie w Wirtualnym Laboratorium Transkrypcji opracowanym w ramach wspomnianego wcześniej projektu SYNAT (http://wlt.synat.pcss.pl/wlt-web/project.xhtml?project=40), a studenci z moich zajęć dokonali transkrypcji kilkunastu stron zgodnie z opracowanymi przeze mnie *ad hoc* zasadami – por. rysunek 3. System okazał się jednak niewygodny i prace te nie są kontynuowane.



Rysunek 3. „Nowy Karakter Polski" w Wirtualnym Laboratorium Transkrypcji

7. Zakończenie

Jesteśmy niestety jeszcze dość daleko od zestawienia pełnego repertuaru znaków piśmiennych dawnej polszczyzny, a jeszcze dalej od ustalenia powszechnie akceptowanych zasad ich kodowania.

Warto tutaj odnotować jednodniową konferencję *Korpusy tekstów dawnych i gwarowych*, która odbyła się 21 listopada 2013 r. w Pracowni Historii Języka Polskiego XVII i XVIII w. Instytutu Języka Polskiego PAN w Warszawie, zorganizowana z inicjatywy Włodzimierza Gruszczyńskiego, pełniącego m.in. funkcję kierownika projektu *Elektroniczny korpus tekstów polskich XVII i XVIII w. (do 1772 r.)* (grant Narodowego Programu Rozwoju Humanistyki). W konferencji wzięli w szczególności udział kierownicy i wykonawcy projektów: *Korpus polszczyzny XVI wieku. Etap I: Digitalizacja źródeł oraz stworzenie narzędzi informatycznych i udostępnienie materiałów testowych korpusu* (grant Narodowego Programu Rozwoju Humanistyki), *Automatyczna analiza fleksyjna tekstów polskich z lat 1830–1918 z uwzględnieniem zmian w odmianie i pisowni 2013–2016* (grant Narodowego Centrum Nauki w ramach konkursu Opus 4), *Fontes mediae et infimae Latinitatis Polonorum. Elektroniczny korpus języka łacińskiego na ziemiach polskich (1000–1550)* (grant Narodowego Programu Rozwoju Humanistyki). Jest nadzieja, że konferencja ta zapoczątkuje współpracę tych zespołów również w kwestii ujednolicenia kodowania tekstów dawnych.

Rozwój korpusów dawnych tekstów polskich prędzej lub później wymusi niezbędną standaryzację.

Bibliografia:

Bernacki L., 1918, *Pierwsza książka polska. Studyum bibliograficzne*, Lwów: Zakład Narodowy im. Ossolińskich. Dostępny w Interecie: http://www.sbc.org.pl/publication/15766.

Bień J. S., 2011, *Podstawowe elementy tekstów elektronicznych*, w: M. Bańko, D. Kopcińska (red.), *Różne formy, różne treści*, Warszawa: Wydział Polonistyki Uniwersytetu Warszawskiego, s. 17–24. Dostępny w Internecie: http://bc.klf.uw.edu.pl/83/.

Bień J. S., 2012a, *Delivering the IMPACT project Polish Ground-Truth texts with Poliqarp for DjVu*. Dostępny w Internecie: http://bc.klf.uw.edu.pl/289/.

Bień J. S., 2012b, *Narzędzia dygitalizacji tekstów na potrzeby badań filologicznych*. Dostępny w Internecie: http://bc.klf.uw.edu.pl/297/.

Bień J. S., 2012c, *Skanowane teksty jako korpusy*, „Prace Filologiczne" 63, s. 25––36. Dostępny w Internecie: http://bc.klf.uw.edu.pl/322/.

Bień J. S., 2014, *The IMPACT project Ground-Truth texts as a DjVu corpus*, „Cognitive Studies/Études Cognitives" 14, s. 75–84. Dostępny w Internecie: http://bc.klf.uw.edu.pl/381/.

Driscoll M. J., 2006, *Levels of transcription*, w: L. Burnard, K. O'Brien O'Keeffe, J. Unsworth (red.), *Electronic Textual Editing*, Modern Language Association of America. Dostępny w Internecie: http:// www.tei-c.org/About/Archive_new/ETE/Preview/driscoll.xml.

Fredell J., Borchers IV Ch., Ilgen T., 2013, *TEI P5 and Special Characters Outside Unicode*, „Journal of the Text Encoding Initiative" 4. Dostępny w Internecie: http://jtei.revues.org/727.

Górski K. et al. (red.), 1955, *Zasady wydawania tekstów staropolskich: projekt*, Wrocław: Zakład im. Ossolińskich.

Haralambous Y., 2007, *Fonts & Encodings. From Advanced Typography to Unicode and Everything in Between*, O'Reilly Media.

Heliński M., Kmieciak M., Parkoła T., 2012, *Report on the comparison of Tesseract and ABBYY FineReader OCR engines*, Tech. rep. Poznań. Dostępny w Internecie: http://lib.psnc.pl/publication/428.

Januszowski J., 1594, *Nowy karakter Polski: z drukarnie Lazarzowey y ortographia polska Iana Kochanowskiego, Ie M. P. Lukasza Gornickiego etc. etc.* Dostępny w Internecie: http://wbl.klf.uw.edu.pl/62/.

Opaliński K., 2007, *Problemy kodowania korpusów historycznych (na przykładzie tekstów XVI-wiecznych)*, w: J. Kamper-Warejko, I. Kaproń-Charzyńska (red.), *Z zagadnień leksykologii i leksykografii języków słowiańskich*, Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, s. 107–114.

Piekarski K., 1936, *Polonia Typographica Saeculi Sedecimi. Zeszyt I. Kasper Hochfeder, Kraków 1503–1506*, wydali Kazimierz Piekarski, Komitet Organizacyjny Zjazdu Bibljotekarzy Polskich w Warszawie. Dostępny w Internecie: http://teksty.klf.uw.edu.pl/40.