Kwartalnik Językoznawczy
DOI: 10.14746/kj.2016.4 / 2017.1.1

Janusz S. Bień

Repertuar znaków piśmiennych – problemy i perspektywy

Wstęp

„Piśmienny" jest przeze mnie rozumiany – za SJPDor – po prostu jako „mający związek z pismem"; celowo używam tutaj tego ogólnego terminu.

Ustalenie repertuaru znaków polskich tekstów – nawet jeśli niektóre z nich będą pomijane lub transkrybowane – jest potrzebne m.in. w rozwijających się pracach nad korpusami dawnych tekstów polskich[1]. W artykule przedstawiam pewne koncepcje metodologiczne i wybrane narzędzia informatyczne, które mogą być pomocne przy realizacji tego zadania.

Czcionki drukarskie

Najbardziej konkretnym punktem wyjścia do badania repertuaru znaków piśmiennych są czcionki drukarskie. Niektóre z nich zachowały się i są dostępne w muzeach drukarstwa, inne znamy tylko ze sporządzonych za ich pomocą druków. Do dotyczących ich badań coraz częściej stosuje się termin *paleotypografia*[2]. W Polsce w użyciu jest termin o zbliżonym znaczeniu, mianowicie *metoda typograficzna [w badaniach nad dawną książką]*. Używa go np. Henryk Bułhak.

---

[1] Por. np. M. Król i in., *Narodowy Korpus Diachroniczny Polszczyzny. Projekt*, „Język Polski" 2019, nr 99 (1), s. 92–101, DOI: 0.31286/JP.99.1.8.

[2] Por. np. J. André, R. Jimenes, *Transcription et codage des imprimés de la Renaissance*, „Revue des Sciences et Technologies de l'Information", Série Document Numérique, 2013, nr 16 (3), DOI: 10.3166/DN.16.3.113-139, https://halshs.archives-ouvertes.fr/halshs-00983575, s. 115–116 [dostęp online: 8.01.2020].

Pierwsze stadium pracy bibliologa-„typografa" polega na uzyskaniu możliwie pełnej znajomości materiału typograficznego poszczególnych tłoczni badanego obszaru drogą gromadzenia, opisu i systematyki zasobów już zidentyfikowanych pod względem pochodzenia (na podstawie druków noszących ten sam adres wydawniczy i wykazujących te same cechy materiału drukarskiego), a następnie na rejestracji zauważonych zmian (przeróbek) i uszkodzeń elementów typograficznych w wyniku ich użytkowania. O ile bowiem sam materiał służy identyfikacji i pozwala przypisać daną pozycję konkretnej oficynie, o tyle zmiany i uszkodzenia elementów, zwłaszcza powiązane z konkretnymi datami druków, stanowią niezwykle ważne, częstokroć jedyne wyznaczniki umożliwiające zabiegi chronologizacyjne[3].

Czcionki dzielono na ślepe, służące do tworzenia odstępów, czyli justunek, i czcionki z oczkiem, które odbijały na papierze znaki odpowiadające kształtowi oczka – kształt ten nazywam *typoglifem*, przez analogię do już istniejącego terminu *glif* (ang. *glyph* – patrz niżej); angielskim odpowiednikiem byłby naturalnie *typoglyph*.

Typoglify czcionek, nawet odlanych z tej samej matrycy, różnią się minimalnie, ale jeśli nie badamy proweniencji i nie przygotowujemy ekspertyzy kryminologicznej, to możemy te różnice zignorować. Typoglify na tym poziomie abstrakcji będziemy nazywać *znakami typograficznymi* (można by je nazywać w skrócie *typoznakami*, w języku angielskim *typographical characters* skracałbym bez wahania jako *typochars*).

Wszechobecność komputerów spowodowała, że czcionki mają teraz postać elektroniczną. Każdemu *znakowi kodowemu* zapisanemu w pamięci komputera przyporządkowują – czasami w dość skomplikowany sposób – *glif*, czyli obraz znaku drukowany lub wyświetlany na odpowiednim urządzeniu. Zdarza się, że czcionki elektroniczne są mniej lub bardziej wiernymi replikami czcionek tradycyjnych.

Wspomniane znaki kodowe to od dłuższego już czasu praktycznie wyłącznie znaki zgodne ze standardem Unicode. Standard ten tylko w niewielkim stopniu i dość niekonsekwentnie uwzględnia potrzeby edycji krytycznych[4] i praktycznie w ogóle nie uwzględnia nowych potrzeb wynikających

[3] H. Bułhak, *Metoda typograficzna w badaniach nad dawną książką. Uwagi i refleksje*, „Biuletyn Poligraficzny" 1977, nr 10 (2), s. 41, http://skryba.inib.uj.edu.pl/~gruca/Teoria%20i%20metodologia%20nauki%20o%20ksiazce/6-Metoda%20typograficzna%20w%20badaniach%20nad%20dawna%20ksiazka.pdf [dostęp online: 8.01.2020].

[4] Por. np. J.S. Bień, *Standard Unicode i język polski*, „Acta Poligraphica" 2019, nr 14, s. 7–28, http://www.cobrpp.com.pl/actapoligraphica/uploads/AP2019_2_Bien.pdf [dostęp online: 8.01.2020]; J.S. Bień, *Traktat Parkosza. Eksperymentalna edycja elektroniczna*, „Poznańskie Studia Polonistyczne. Seria Językoznawcza" 2019, nr 26 (1), s. 27–69, http://pressto.amu.edu.pl/index.php/pspsj/article/view/19852 [dostęp online: 8.01.2020].

z rozwoju technik automatycznego rozpoznawania znaków (OCR – *Optical Character Recognition*), które coraz częściej stosuje się z lepszym lub gorszym skutkiem do tekstów dawnych. Nie ma jednak dla niego alternatywy. Związek znaków kodowych ze znakami typograficznymi i znakami piśmiennymi pozostaje zawiły i nieoczywisty. Jest to odrębny temat, który tylko częściowo omówiłem w dwóch artykułach[5].

Warto dodać, że wykształciły się dwie tradycje składu ręcznego, nazwane przeze mnie umownie anglosaską i kontynentalną, różniące się m.in. terminologią. Terminologia anglosaska docierała do Polski razem z urządzeniami do mechanizacji składu – linotypami i monotypami, ale posługiwanie się nią ograniczone było do personelu drukarni. Jednak razem z komputerami osobistymi trafiły do Polski programy składania tekstów dostosowane do anglosaskich zwyczajów typograficznych i stosujące w dokumentacji terminologię anglosaską, z którą tłumacze nie zawsze potrafili sobie poradzić. Pojęciem niemającym wcześniej odpowiednika w terminologii polskiej jest *font*, czyli czcionki pochodzące z jednego „odlewu"; puryści tłumaczą to słowo jako 'czcionki', wprowadzając niepotrzebną wieloznaczność. Należę do tych, którzy uważają *font* za równie dobre polskie słowo jak np. *fontanna*.

Polonia Typographica

W 1920 roku Ludwik Bernacki (m.in. wieloletni dyrektor Ossolineum) opublikował koncepcję *Monumenta Poloniae Typographica XV et XVI Saeculorum*[6]. *Monumenta* miały w szczególności zawierać „atlas zupełnego zasobu typograficznego oficyn ówczesnych na ziemiach Rzeczypospolitej"[7]. Tak przynajmniej uważał Kazimierz Piekarski (ja nie znalazłem tekstu Bernackiego zawierającego taki postulat), który podjął się częściowej jego realizacji w pracy *Pierwsza drukarnia Florjana Unglera 1510–1516: chronologia druków i zasobu typograficznego*[8], a następnie inicjując serię *Polonia*

[5] Por. *ibidem*.

[6] L. Bernacki, *Monumenta typographica Poloniae XV et XVI Saeculorum,* „Exlibris. Pismo poświęcone bibljofilstwu polskiemu" 1920, z. III, s. 89–90, http://kpbc.ukw.edu.pl/publication/11872 [dostęp online: 8.01.2020].

[7] K. Piekarski, *Kasper Hochfeder, Kraków 1503–1505*, *Polonia typographica saeculi sedecimi: Zbiór podobizn zasobu drukarskiego tłoczni polskich XVI stulecia*, z. 1, Warszawa 1936, s. 5, http://ebuw.uw.edu.pl/publication/161144 [dostęp online: 8.01.2020].

[8] K. Piekarski, *Pierwsza drukarnia Florjana Unglera 1510–1516: chronologia druków i zasobu typograficznego*, Kraków 1926, http://kpbc.umk.pl/publication/17339 [dostęp online: 8.01.2020].

typographica saeculi sedecimi: Zbiór podobizn zasobu drukarskiego tłoczni polskich XVI stulecia, której pierwszy „zeszyt" opublikował, częściowo własnym sumptem, w 1936 roku[9].

W ramach serii opublikowano 12 „zeszytów", będących w istocie teczkami o rozmiarach ok. 41 na 18 cm. Każda teczka zawiera pewną liczbę luźnych jednostronnych planszy, na których znajdują się tablice, oraz mniejszego formatu broszurę z tekstem opisowym. Łączna liczba tablic wynosi 599. Wiele z nich przedstawia drzeworyty używane jako ilustracje lub ozdobniki, bogato reprezentowane pozostają również ozdobne inicjały, ale oczywiście są także tablice przedstawiające czcionki tekstowe.

Przed II wojną światową wydano tylko dwa zeszyty. Inicjatorką wznowienia serii po wojnie była Alodia Kawecka-Gryczowa, która była również redaktorem serii (także po przejściu na emeryturę).

Zeszyt 1 miał drugie wydanie przejrzane i poprawione w 1968 roku, opracowała je Maria Błońska. Zeszyt 2, wydany przez Piekarskiego w 1937, również miał w 1963 roku drugie wydanie przejrzane, które opracowała Helena Kapełuś (przygotowała ona również uzupełnienia i poprawki do zeszytu 4, który ukazał się rok wcześniej). W 1959 opublikowano zeszyt 3, który opracował Bułhak, w 1962 roku zeszyt 4, który opracowała Kapełuś, a w latach 1964, 1966 i 1970 zeszyty 5, 6 i 7, które opracował Bułhak (zawarte w zeszycie 7 indeksy przygotowała Anna Wolińska). Zeszyt 8, wydany w 1972 roku, opracowała Paulina Buchwald-Pelcowa. Zeszyt 9 wydany w 1972, zeszyt 10 wydany w 1975 i zeszyt 11 wydany w 1981 roku opracowała Kawecka-Gryczowa. Zeszyt 12, wydany w 1981, opracował Bułhak.

Piekarski zmarł w 1944 roku, co oznacza, że jego publikacje należą już do domeny publicznej. Dygitalizacja zeszytu 1 jest dostępna w całości w e-BUW (http://ebuw.uw.edu.pl/publication/161144) (do niedawna w formacie DjVu, obecnie m.in. w formacie PDF). Biblioteka cyfrowa Polona udostępnia z nieznanych powodów tylko tabele zeszytu 1 i 2[10]. Status prawny pozostałych zeszytów jest niejasny. Jeśli traktować je w całości jako wydania krytyczne, to prawa do nich wygasły 30 lat po publikacji; jeśli jednak części opisowe traktować jako normalne utwory, to prawa do nich wygasają dopiero 70 lat od śmierci autora. Nie sądzę, żeby autorzy byli przeciwni szerokiemu udostępnieniu zeszytów w Internecie, ale przy drugiej interpretacji należałoby od nich lub ich spadkobierców uzyskać jawne zgody (Kawecka-Gryczowa zmarła w 1990, Kapełuś w 1999, a Błońska w 2008 roku).

[9] Por. wstęp do: K. Piekarski, *Kasper Hochfeder…*

[10] https://polona.pl/item/54430363, https://polona.pl/item/54430364.

Udostępnienie serii w Internecie nie powinno się ograniczać do prezentacji skanów. Dla dygitalizacji w e-BUW został wykonany tzw. brudny OCR, czyli automatyczne rozpoznanie znaków; w części opisowej jego jakość wydaje się akceptowalna. Możliwe są również liczne inne ulepszenia wskanowanego tekstu, jak dodanie adnotacji ukazujących się po najechaniu myszą na odpowiedni fragment, uzupełnienie indeksów o odpowiednie hiperlinki itp. Oczywiście w XXI wieku kontynuowanie prac nad *Polonia Typographica* według koncepcji sprzed prawie 100 lat nie miałoby sensu. Do analizy typograficznej druków należałoby użyć odpowiednich narzędzi informatycznych, a wyniki prezentować w formie elektronicznej w stosownym formacie.

Warto dodać, że analiza konkretnych druków nie jest jedyną metodą, ponieważ aż do czasów współczesnych drukarnie przygotowywały wzorniki czcionek (przeważnie dla swoich klientów, ale np. znajdujący się w BUW niedatowany wzornik[11] był *DO PRZESWIETNEY POLICYI PODANY* [pisownia oryginalna]). Trudno powiedzieć, jak dużo z nich się zachowało. Zdygitalizowanych wzorników jest chyba mało, ze względu na nieoczywisty sposób katalogowania niełatwo je wyszukać. Moim zdaniem wzorniki zasługują na większą uwagę, w szczególności na wyczerpującą dygitalizację i jakiś zbiorczy katalog.

Koncepcje metodologiczne

Możliwe podejścia do problemu przedstawię na przykładzie koncepcji dwóch projektów francuskich *Cassetin* i *PICA*.

Projekt *Cassetin* zaproponował informatyk (chyba już wtedy emerytowany) Jacques André[12] (słowo *cassetin* oznacza po francusku rodzaj kaszty drukarskiej, ale jako nazwa projektu ma być traktowane jako skrót od *CASSE Type encodINg* – w bardzo swobodnym tłumaczeniu ‘kodowanie czcionek drukarskich’). Zadania projektu zostały sformułowane następująco[13]:

– sporządzenie inwentarza wszystkich czcionek używanych w tekstach drukowanych w językach europejskich,

[11] Według Marii Judy pochodzący z 1790 roku; por. *eadem*, *Pismo drukowane w Polsce XV-XVIII wieku*, Lublin 2001, s. 347, http://dlibra.umcs.lublin.pl/publication/597 [dostęp online: 8.01.2020].

[12] J. André, *Numérisation et codage des caractères de livres anciens*, „Document numérique” 2003, vol. 7, nr 3, s. 127–142, http://dn.revuesonline.com/article.jsp?articleId=2325 [dostęp online: 8.01.2020]; *idem*, *The Cassetin Project – Towards an Inventory of Ancient Types and the Relate Standardised Encoding*, „TUGboat” 2003, vol. 24, nr 3, s. 314–318, http://www.tug.org/TUGboat/tb24-3/andre.pdf [dostęp online: 8.01.2020].

[13] J. André, *The Cassetin Project*, s. 317–318.

– nazwanie czcionek i ich zakodowanie, tj. przypisanie nazwie jakiejś liczby należącej do tzw. obszaru użytku prywatnego standardu Unicode,
– stworzenie eksperymentalnego fontu.

Propozycja ta – merytorycznie sensowna, choć nierealistycznie ambitna – była zaproszeniem do współpracy m.in. w celu wspólnego wystąpienia o środki finansowe na realizację tego przedsięwzięcia. To się najwyraźniej nie udało.

Jako kontynuację projektu *Cassetin* potraktowałbym publikację Andrégo i Jimenesa *Transcription et codage des imprimés de la Renaissance*[14], choć autorzy posługują się nazwą PICA (*projet d'inventaire des caractères anciens* – projekt inwentarza dawnych znaków). Artykuł przypomina rozróżnienie rodzajów transkrypcji wprowadzone przez redaktorów elektronicznej edycji *Opowieści kanterberyjskich* Chaucera[15]:

– transkrypcja graficzna (celowość tego wyróżnienia jest kwestionowana),
– transkrypcja grafetyczna, która zachowuje np. długie i krótkie *s*,
– transkrypcja grafemiczna,
– transkrypcja znormalizowana, która obejmuje w szczególności rozwinięcia skrótów.

Oryginalną koncepcją autorów jest wprowadzenie pojęcia bazującego na kilkusetletniej tradycji klasyfikowania czcionek przez drukarzy, w której to klasyfikacji jednostką była zawartość jednej przegródki kaszty (w Polsce był w użyciu termin *króbka*). Pojęciu temu odpowiada termin *typem* (fr. *typème*). Dla różnych stopni (wielkości) i krojów pisma przeznaczone były odrębne kaszty, ale o tym samym układzie, więc stopień (wielkość) i krój nie są własnością typemu. Konsekwencją tej decyzji jest wprowadzenie typemicznego (fr. *typémique*) poziomu transkrypcji tekstu, poprzedzającego poziom grafetyczny, innymi słowy zastępującego poziom graficzny wspomniany wcześniej.

Wprowadzają również ilustrowane konkretnymi przykładami pojęcie *znaku symulowanego*, np. brakująca czcionka *k* mogła być reprezentowana przez *lz* (polskim przykładem może być zapis *w* jako *rv*). W transkrypcji typemicznej są one reprezentowane wiernie jako *lz* (odpowiednio *rv*). Mówiąc

[14] J. André, R. Jimenes, *op.cit.*
[15] P. Robinson, E. Solopova, *Guidelines for Transcription of the Manuscripts of The Wife of Bath's Prologue*, 2006, http://server30087.uk2net.com/canterburytalesproject.com/pubs/transguide-MI.pdf [dostęp online: 8.01.2020].

bardziej technicznie, transkrypcja taka to tekst elektroniczny zgodny ze standardem Unicode, ale z dwoma zastrzeżeniami. Pierwsze jest oczywiste – to stosowanie znaków z tzw. obszaru użytku prywatnego, zarówno tych zdefiniowanych w rekomendacjach MUFI (*Medieval Unicode Font Initiative*[16]), jak i tych zaproponowanych przez autorów, np. LETTRE MINUSCULE LATINE E AVEC PARAPHE (mała litera łacińska *e* z parafą – z ilustracji wynika, że chodzi o literę *e* z diakrytem podobnym do apostrofu) i znaki do reprezentacji „incydentów typograficznych" (pomyłek lub celowych odstępstw od zasad): CARACTÈRE PIED-EN-HAUT (czcionka do góry nogami), ERREUR DE POSITION (błędne położenie czcionki).

Drugie zastrzeżenie wiąże się z faktem, że standard Unicode przewiduje odrębne kody dla znaków nie zawsze odróżnialnych w druku, np. LATIN CAPITAL LETTER RUM ROTUNDA (skrót sylaby *rum* w formie przekreślonej litery *r rotunda*), JUPITER (astrologiczny symbol Jowisza), ALCHEMICAL SYMBOL FOR TIN ORE (alchemiczny symbol rudy cyny). W takich sytuacjach na potrzeby transkrypcji typemicznej zostanie arbitralnie wybrany jeden z tych znaków, a oryginalne rozróżnienia semantyczne pojawią się dopiero na poziomie grafetycznym.

Jak widać, postulowany przeze mnie *znak typograficzny* jest bardziej konkretny niż *typem*, bo jest określonej wielkości i kroju. Z drugiej strony postulowany przeze mnie w innych publikacjach[17] *tekstel* jest pojęciem bardziej abstrakcyjnym od *typemu*, można go chyba przypisać do poziomu grafetycznego.

Projekt PICA miał bazować na systemie BaTyR (*Base de Typographie de la Renaissance* – w swobodnym tłumaczeniu baza typografii renesansowej[18]) – którego wstępna wersja została udostępniona w marcu 2014 roku. Miały być również wykorzystywane narzędzia projektu PaRADIIT (*Pattern Redundancy Analysis for Document Image Indexation & Transcription* – w swobodnym tłumaczeniu *Analiza redundancji form na potrzeby indeksowania i transkrypcji dokumentów w postaci graficznej*[19]). Narzędzia te rzeczywiście wyglądały obiecująco, ale mimo kilku prób nie udało mi się ich sensownie użyć na moim komputerze (autorzy nie byli chętni do udzielenia pomocy – może byli zbyt zajęci). Projekt PaRADIIT chyba nie jest

[16] https://mufi.info/.

[17] Zob. np. J.S. Bień, *Problemy kodowania znaków w korpusach historycznych*, w: *Semantyka a konfrontacja językowa*, red. D. Roszko i J. Satoła-Staśkowiak, t. 5, Warszawa 2016, s. 67–76, https://www.researchgate.net/publication/338448462 [dostęp online: 8.01.2020].

[18] http://www.bvh.univ-tours.fr/batyr/beta/index.php [dostęp online: 22.01.2020].

[19] https://sites.google.com/site/paradiitproject/ [dostęp online: 22.01.2020].

kontynuowany, a projekt PICA pozostał w fazie koncepcyjnej. Tym niemniej różne wyniki cząstkowe, jak baza BaTyR i proponowany aparat pojęciowy, zasługują moim zdaniem na uwagę.

Narzędzia informatyczne

Nie jest moim celem dokonanie tutaj szerokiego przeglądu narzędzi informatycznych, które z grubsza można podzielić na dwie grupy: do opisywania pojedynczych znaków na potrzeby trenowania programów do OCR i do transkrybowania słów, wierszy i całych tekstów; transkrybowanie może być celem samym w sobie lub środkiem do trenowania lub oceniania programów rozpoznawania struktury dokumentów[20].

W swoim czasie eksperymentowałem z wieloma programami, które wydawały mi się obiecujące. Najlepsze wrażenie zrobił na mnie system Transkribus[21], z którego od czasu do czasu nadal korzystam. Niestety po zakończeniu grantu projekt musi się samofinansować, w związku z czym bardziej intensywne użycie jest od 2019 roku odpłatne. Eksperymentowałem też z programami Gamera[22], PoCoTo[23] i Glyph Miner[24]. Niestety miałem z nimi różne problemy i nie udało mi się wdrożyć żadnego z nich do użytku. Zainteresowani Czytelnicy mogą sprawdzić samodzielnie ich aktualne wersje.

W konsekwencji tych problemów tutaj ograniczę się do przedstawienia dwóch własnych koncepcji, które niestety zostały zrealizowane tylko w formie bardzo wstępnych prototypów.

---

[20] O tej grupie programów pisała Małgorzata Kowalska, *Transkrypcja tekstów w środowisku elektronicznym. Przegląd wybranych narzędzi*, „Sztuka Edycji" 2016, nr 10 (2), s. 65–74, https://apcz.umk.pl/czasopisma/index.php/sztukaedycji/article/view/SE.2016.020 [dostęp online: 8.01.2020].

[21] https://transkribus.eu/ [dostęp online: 22.02.2020]; zob. także: J.S. Bień, *Traktat Parkosza…*, s. 35.

[22] C. Dalitz, R. Baston, *Optical Character Recognition with the Gamera Framework*, w: *Document Image Analysis with the Gamera Framework*, red. C. Dalitz, t. 8, *Schriftenreihe des Fachbereichs Elektrotechnik und Informatik*, 2009, s. 53–65, https://www.researchgate.net/publication/281267308 [dostęp online: 8.01.2020].

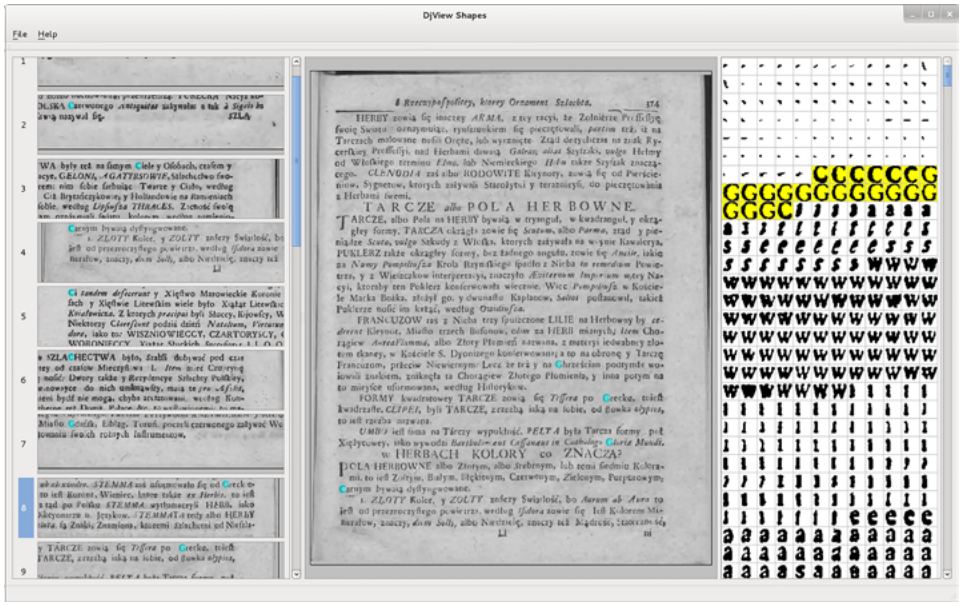
[23] F. Fink, U. Springmann, *Postcorrection Tool (PoCoTo) Manual*, 2017, s. 38, https://github.com/cisocrgroup/Resources/blob/master/manuals/pocoto-manual.pdf [dostęp online: 8.01.2020].

[24] B. Budig, T.C. van Dijk, F. Kirchner, *Glyph miner: A system for efficiently extracting glyphs from early prints in the context of OCR*, w: 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2016, s. 31–34, http://www1.informatik.uni-wuerzburg.de/fileadmin/10030100/Presentation-JCDL2016.pdf [dostęp online: 8.01.2020].

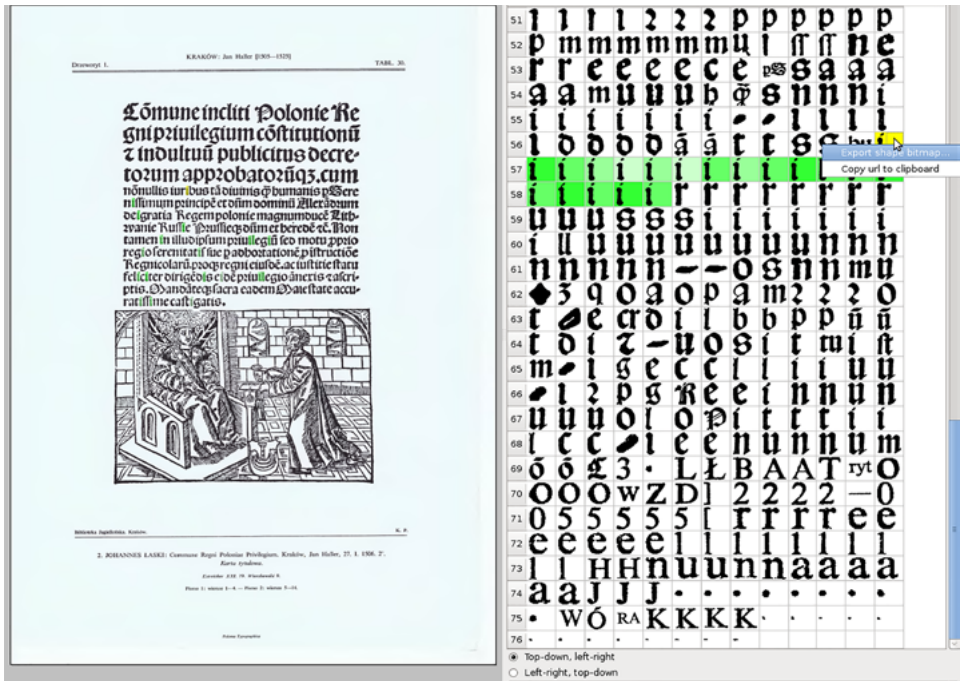
Obecnie format DjVu jest w odwrocie (nie jest obsługiwany przez przeglądarki WWW), ale wcześniej większość publikacji polskich bibliotek cyfrowych była przechowywana i udostępniana właśnie w tym formacie i teraz jest być może podobnie (aktualne dane statystyczne nie są dostępne).

Wysoki stopień kompresji zapewniany przez format DjVu bierze się stąd, że obraz zostaje rozdzielony na tło i zadruk. Następnie wykorzystuje się fakt, że zadruk zawiera powtarzające się odbicia czcionek. Odbywa się to w ten sposób, że jeden z obrazów danej czcionki jest zapisany w całości, a dla innych zapisuje się tylko różnice, z reguły niewielkie, między obrazem podstawowym a danym odbiciem czcionki. Zapis tych różnic ma formę drzewa nazywanego – całkiem logicznie – słownikiem kształtów. Zapisane w ten sposób kształty stanowią z reguły dość dobre przybliżenie repertuaru znaków. Jest to tylko przybliżenie, ponieważ kształty w słownikach są spójne, co powoduje rozdzielenie dwuczęściowych liter typu *c* z kreską, *z* z kropką, a także litery *i*. Wydaje się jednak, że z praktycznego punktu widzenia nawet takie przybliżenie może być pożyteczne.

Standardowa przeglądarka djview4 została zmodyfikowana w taki sposób, że możliwe jest wygodne przeglądanie kształtów; modyfikacja ta została nazwana djview-shapes – patrz il. 1 (kolory wskazują na poziom drzewa kształtów podobnych).



Il. 1: djview-shapes: wykaz wystąpień kształtów podobnych do litery *C*



Il. 2: djview-shapes: eksport wybranego kształtu

Podstawowe funkcje programu są następujące:

– pojedyncze kliknięcie lewym klawiszem myszy na kształt w bocznym panelu przełącza między wybraniem pojedynczego kształtu a drzewem kształtów, dla którego jest on korzeniem; wystąpienia kształtu lub kształtów są wyświetlane w panelu konkordancji z lewej strony,

– pojedyncze kliknięcie prawym klawiszem myszy na kształt w bocznym panelu eksportuje wybrane kształty w formie plików graficznych – por. il. 2,

– pojedyncze kliknięcie lewym klawiszem w panelu głównym wybiera odpowiedni kształt w bocznym panelu,

– podwójne kliknięcie lewym klawiszem myszy na panel konkordancji wyświetla odpowiednią stronę w oknie głównym,

– kliknięcie środkowym klawiszem myszy na panel konkordancji uruchamia nowy egzemplarz djview4 wyświetlający odpowiednią stronę,

– wciśnięcie lewego klawisza myszy pozwala pozycjonować tekst w oknie,

– wciśnięcie klawisza Ctrl pozwala wykorzystać rolkę myszy do skalowania tekstu w oknie.

Kształty w prawym panelu są wyświetlane dla liczby stron wskazanej w ustawieniach.

Ponieważ zaletą formatu było szybkie udostępnianie dokumentów przez Internet, typowy dokument zawiera słowniki kształtów tylko dla poszczególnych stron lub dla ich niewielkiej liczby – inaczej niepotrzebnie pobierano by dane nieistotne dla wyświetlanej strony. Podanie dużej liczby stron w programie djview-shapes może wymagać połączenia kilku takich słowników. W prototypie okazało się to bardzo czasochłonne – na optymalizację tych obliczeń (np. na wykonywanie ich w osobnym programie przed rozpoczęciem przeglądania kształtów) zabrakło czasu i funduszy w grancie, w ramach którego został zrealizowany program (N N519 384036 *Narzędzia dygitalizacji tekstów na potrzeby badań filologicznych*, 2009–2012).

Prototyp programu zrealizował Grzegorz Chimosz, wersję ostateczną prawie od nowa wykonał Michał Rudolf. Program istnieje tylko w wersji dla systemu Linux (nie było zapotrzebowania, a także czasu i funduszy na wersję dla MS Windows). Gotowy do użytku program znajduje się na maszynie wirtualnej sid4ocr zdeponowanej w repozytorium CLARIN-PL (*Common Language Resources & Technology Infrastructure*) pod adresem https://clarin-pl.eu/dspace/handle/11321/469. Tekst źródłowy programu znajduje się w repozytorium https://bitbucket.org/mrudolf/djview-shapes.

Warto dodać, że opracowanie narzędzia do tworzenia inwentarza znaków było jednym z oficjalnych celów europejskiego projektu IMPACT (*IMProving ACcess to Texts*[25]), ale prace te zostały przerwane właściwie bez wyjaśnienia. Uzyskane wyniki zostały udostępnione publicznie na wolnej licencji[26], od 2012 roku nie są jednak rozwijane. Niestety oprogramowanie to z założenia jako danych wejściowych wymaga wyników uzyskanych dzięki komercyjnym programom używanym w projekcie, lecz niedostępnych dla zwykłego użytkownika. W rezultacie przydatność tego programu jest znikoma.

Zadanie stworzenia programu określającego repertuar konkretnego dzieła nadal czeka na chętnego programistę lub sponsora.

Program djview-shapes pozwala tylko zorientować się w repertuarze znaków, na potrzeby transkrypcji lub trenowania programów rozpoznawania znaków potrzebne jest narzędzie do etykietowania zidentyfikowanych typoglifów. Próbę stworzenia takiego programu podjęto we wspomnianym grancie *Narzędzia dygitalizacji tekstów na potrzeby badań filologicznych*,

[25] Zob. np. M. Guy, *IMPACT Final Conference 2011*, „Ariadne" 2012, nr 68, http://www.ariadne.ac.uk/issue/68/impact-rpt/ [dostęp online: 8.01.2020].

[26] https://github.com/impactcentre/inventory-extraction [dostęp online: 22.01.2020].

niestety wskutek nieefektywnej implementacji prototyp okazał się zbyt wolny w działaniu, aby mógł być w jakikolwiek sposób użyteczny (zainteresowane osoby mogą ocenić to samodzielnie, program jest dostępny publicznie razem z innymi wynikami projektu – patrz https://bitbucket.org/jsbien/ndt). Tym niemniej pomysł wydaje się wart przedstawienia.

Chodzi przede wszystkim o koncepcję pracy w układzie klient–serwer (obecnie zaczyna to być dość powszechne, tak działa np. wspomniany wcześniej Transkribus). Informacje o znakach miały być przechowywane w bazie danych na serwerze, która mogłaby być wykorzystywana zarówno lokalnie, jak i zdalnie.

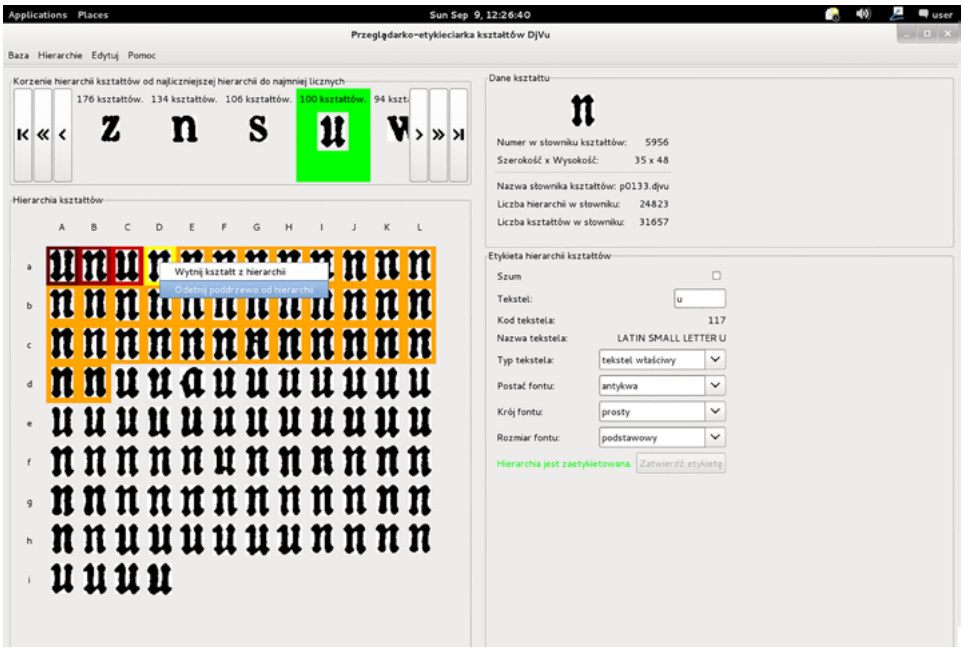
Program kliencki służył do etykietowania całych hierarchii kształtów. Przypisana informacja składała się z następujących własności:

– umowna nazwa (w miarę możliwości zgodna ze standardem Unicode),
– kod liczbowy (w miarę możliwości zgodny ze standardem Unicode),
– typ tekstela: *tekstel właściwy*, *mikrotekstel* (teraz powiedziałbym raczej *mikroglif*) – fragment znaku, np. sama kropka od litery *i*, *makrotekstel* (teraz powiedziałbym raczej *makroglif*) – znaki zlepione przez nadmiar farby drukarskiej,
– postać fontu (teraz powiedziałbym raczej *typ pisma*) – chodziło przede wszystkim o odróżnienie gotyku od łacinki,
– krój fontu: np. prosty, pochyły, wytłuszczony,
– rozmiar fontu: np. podstawowy, tytułowy.

Własności te mogą wydawać się oczywiste, ale wiele programów do transkrypcji lub trenowania OCR uwzględnia tylko niektóre z nich, np. brak rozróżnienia między gotykiem i antykwą w korpusie IMPACT[27] jest konsekwencją tego właśnie faktu.

Ilustracja 3 pokazuje jeden z etapów etykietowania drzewa kształtów litery *u*, cały proces ilustruje łącznie 14 zrzutów ekranu dostępnych na witrynie grantu https://bitbucket.org/jsbien/ndt/wiki/Home. Hierarchie kształtów są generowane odrębnym programem, teoretycznie nie jest więc konieczne ograniczanie się do dokumentów DjVu.

[27] J.S. Bień, *The IMPACT project Polish Ground-Truth texts as a DjVu corpus*, „Cognitive Studies | Études Cognitives" 2014, nr 14, s. 75–84, https://ispan.waw.pl/journals/index.php/cs-ec/article/view/cs.2014.008 [dostęp online: 8.01.2020].



Il. 3: Etykietowanie: przygotowanie do oddzielenia kształtów *n* od drzewa kształtów *u*

Uwagi końcowe

Współczesna wersja *Polonia Typographica* powinna mieć moim zdaniem postać systemu komputerowego w postaci serwera obsługującego odpowiednią bazę danych, dostępną na co najmniej dwa sposoby. Pierwszy to witryna internetowa pozwalająca na wprowadzanie obrazów znaków w sposób podobny jak ma to miejsce przy szukaniu obrazem w wyszukiwarce Google; prezentacja wyników mogłaby być inspirowana wyszukiwarką znaków standardu Unicode dostępną pod adresem http://shapecatcher.com/ (trafienia są uporządkowane według stopnia podobieństwa wyszukiwanego obiektu do dostępnych wzorców, użytkownik ma możliwość przekazania swojej oceny wyniku itp.). Drugi sposób dostępu to tzw. API (*Application Programming Interface* – interfejs programistyczny aplikacji) pozwalający na tworzenie programów klienckich umożliwiających automatyzację pewnych operacji lub wprowadzających inne ułatwienia.

Umieszczona na serwerze baza danych o znakach mogłaby komasować informacje dotyczące różnych druków i pochodzące od różnych użytkowników, zarówno instytucjonalnych, jak i hobbystów (jest to tzw. *crowdsourcing*), bez ograniczeń co do ich geograficznej lokalizacji – możliwa byłaby zatem nawet współpraca międzynarodowa.

Z informatycznego punktu widzenia zadanie nie jest szczególnie trudne, ponieważ algorytmy są znane i odpowiednie narzędzia pozostają dostępne. Zasadniczy problem to znalezienie form organizacyjnych, które pozwoliłyby nie tylko na stworzenie systemu, ale i na jego niezakłócone funkcjonowanie tak długo, jak byłoby to celowe (czyli np. do czasu zastąpienia przez inny, lepszy system).

Wspomniane repozytoria wyników projektu *Narzędzia dygitalizacji tekstów* nie są już przeze mnie rozwijane, ale będę się starał, aby były dostępne w Internecie przez rok od publikacji niniejszego artykułu, potem zostaną zlikwidowane. Zainteresowane osoby proszone są więc o ich skopiowanie lub zgłoszenie się do mnie w celu przejęcia administracji repozytoriami.

Podziękowania

Niniejszy artykuł został przygotowany – jak wszystkie moje artykuły – za pomocą systemu TeX (konkretnie XeLaTeX). Na życzenie Redakcji został on skonwertowany do formatu Worda (konkretnie do akceptowanego przez Worda formatu ODT) za pomocą programu `make4ht`; w dokonaniu konwersji pomógł mi istotnie autor programu, Michal Hoftich – jestem mu za to bardzo wdzięczny.

Bibliografia

André J., *Numérisation et codage des caractères de livres anciens*, „Document numérique" 2003, vol. 7, nr 3, s. 127–142, http://dn.revuesonline.com/article.jsp?articleId=2325 [dostęp online: 8.01.2020].

André J., *The Cassetin Project – Towards an Inventory of Ancient Types and the Relate Standardised Encoding*, „UGboat" 2003, vol. 24, nr 3, s. 314–318, http://www.tug.org/TUGboat/tb24-3/andre.pdf [dostęp online: 8.01.2020].

André J., Jimenes R., *Transcription et codage des imprimés de la Renaissance*, „Revue des Sciences et Technologies de l'Information", – *Série Document Numérique* 2013, vol. 16, nr 3, s. 113–139, https://halshs.archives-ouvertes.fr/halshs-00983575 [dostęp online: 8.01.2020].

Bernacki L., *Monumenta typographica Poloniae XV et XVI Saeculorum*, „Exlibris. Pismo poświęcone bibljofilstwu polskiemu" 1920, z. III, s. 89–90, http://kpbc.ukw.edu.pl/publication/11872 [dostęp online: 8.01.2020].

Bień J.S., *The IMPACT project Polish Ground-Truth texts as a DjVu corpus*, „Cognitive Studies Études Cognitives" 2014, nr 14, s. 75–84, https://ispan.waw.pl/journals/index.php/cs-ec/article/view/cs.2014.008 [dostęp online: 8.01.2020].

Bień J.S., *Problemy kodowania znaków w korpusach historycznych*, w: *Semantyka a konfrontacja językowa*, t. 5, red. D. Roszko i J. Satoła-Staśkowiak, Warszawa 2016, s. 67–76, https://www.researchgate.net/publication/338448462 [dostęp online: 8.01.2020].

Bień J.S., *Standard Unicode i język polski*, „Acta Poligraphica" 2019, nr 14, s. 7–28, http://www.cobrpp.com.pl/actapoligraphica/uploads/AP2019_2_Bien.pdf [dostęp online: 8.01.2020].

Bień J.S., *Traktat Parkosza. Eksperymentalna edycja elektroniczna*, „Poznańskie Studia Polonistyczne. Seria Językoznawcza" 2019, t. 26. nr 1, s. 27–69, http://pressto.amu.edu.pl/index.php/pspsj/article/view/19852 [dostęp online: 8.01.2020].

Budig B., van Dijk T.C., Kirchner F., *Glyph miner: A system for efficiently extracting glyphs from early prints in the context of OCR*, w: 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL) 2016, s. 31–34, http://www1.informatik.uni-wuerzburg.de/fileadmin/10030100/Presentation-JCDL2016.pdf [dostęp online: 8.01.2020].

Bułhak H., *Metoda typograficzna w badaniach nad dawną książką. Uwagi i refleksje*, „Biuletyn Poligraficzny" 1977, vol. 2, nr 10, s. 37–52,

http://skryba.inib.uj.edu.pl/~gruca/Teoria%20i%20metodologia%20nauki%20o%20ksiazce/6-Metoda%20typograficzna%20w%20badaniach%20nad%20dawna%20ksiazka.pdf [dostęp online: 8.01.2020].

Dalitz Ch., Baston R., *Optical Character Recognition with the Gamera Framework*, w: *Document Image Analysis with the Gamera Framework*, red. Ch. Dalitz, *Schriftenreihe des Fachbereichs Elektrotechnik und Informatik*, vol. 8, 2009, s. 53–65, https://www.researchgate.net/publication/281267308 [dostęp online: 8.01.2020].

Fink F., Springmann U., *Postcorrection Tool (PoCoTo) Manual*, 2017, s. 38, https://github.com/cisocrgroup/Resources/blob/master/manuals/pocoto-manual.pdf [dostęp online: 8.01.2020].

Guy M., *IMPACT Final Conference 2011*, „Ariadne" 2012, nr 68, http://www.ariadne.ac.uk/issue/68/impact-rpt/ [dostęp online: 8.01.2020].

Juda M., *Pismo drukowane w Polsce XV-XVIII wieku*, Lublin 2001, http://dlibra.umcs.lublin.pl/publication/597 [dostęp online: 8.01.2020].

Kowalska M., *Transkrypcja tekstów w środowisku elektronicznym. Przegląd wybranych narzędzi*, „Sztuka Edycji" 2016, vol. 10, nr 2, s. 65–74, https://apcz.umk.pl/czasopisma/index.php/sztukaedycji/article/view/SE.2016.020 [dostęp online: 8.01.2020].

Król M. i in., *Narodowy Korpus Diachroniczny Polszczyzny. Projekt*, „Język Polski" 2019, vol. 99, nr 1, s. 92–101.

Piekarski K., *Pierwsza drukarnia Florjana Unglera 1510–1516: chronologia druków i zasobu typograficznego*, Kraków 1926, http://kpbc.umk.pl/publication/17339 [dostęp online: 8.01.2020].

Piekarski K., *Kasper Hochfeder*, *Kraków 1503–1505*, *Polonia typographica saeculi sedecimi: Zbiór podobizn zasobu drukarskiego tłoczni polskich XVI stulecia*, z. 1, Warszawa 1936, http://ebuw.uw.edu.pl/publication/161144 [dostęp online: 8.01.2020].

Robinson P., Solopova E., *Guidelines for Transcription of the Manuscripts of The Wife of Bath's Prologue*, 2006, http://server30087.uk2net.com/canterburytalesproject.com/pubs/transguide-MI.pdf [dostęp online: 8.01.2020].

Towards an electronic repertoire of basic text elements

SUMMARY

The Unicode standard notions of characters nad glyphs are not intuitive and insufficient for some applications, so there is a need for other notions. Some of them are presented in the paper. Some computer tools for text transcription are also presented. Earlier Polish research in the domain now usually called paleotypography is briefly discussed.

Key words: paleotypography, Unicode, character, glyph, Polish language, Polonia Typographica.

O Autorze

Prof. dr hab. Janusz S. Bień – emeryt (wcześniej m.in. Katedra Lingwistyki Formalnej Wydziału Neofilologii Uniwersytetu Warszawskiego); informatyk i lingwista (z wykształcenia matematyk);
aktualne zainteresowania: dygitalizacja dawnych tekstów polskich, historia pisowni polskiej.
E-mail: jsbien@uw.edu.pl
Witryna: https://sites.google.com/view/jsbien/