Język, Komunikacja, Informacja
I. Koutny, P. Nowak (red.)
Language, Communication, Information
I. Koutny, P. Nowak (eds.)
5/2010–2011: 123–134

Rafał Lewandowski

Społeczna korekta post-OCR w bibliotekach cyfrowych

Abstract (Social Proofreading post-OCR in Digital Libraries). Increasing numbers of libraries are converting their collections to digital form. The digital images are obtained through a process of scanning, and are searched based on metadata which is entered into the system. To enable full-text searching, it is necessary to make use of Optical Character Recognition (OCR) technology. Unfortunately, automatic recognition of text in historical publications is very difficult. This is due to several factors, including low quality of the input due to imprecise printing (smudged characters, duplicated letters, small font size), and the thinness of the paper, which causes the reverse side to show through during scanning. Poor storage conditions (humidity) cause the paper to wrinkle. Improper storage also means that the text is not scanned in a straight line. One of the basic methods used to support the text recognition process is dictionary-based correction (often in real time). More and more often, electronic libraries are making use of public collaborative OCR text correction. This offers a high level of accuracy at low cost.

Abstrakt. Biblioteki coraz częściej dokonują konwersji swoich zbiorów do postaci cyfrowej. Otrzymywane w procesie skanowania obrazy rastrowe są przechowywane na serwerze, a ich wyszukiwanie odbywa się dzięki wprowadzanym do systemu metadanym. Aby umożliwić przeszukiwanie pełnotekstowe niezbędne jest wykorzystanie technologii OCR (ang. Optical Character Recognition). Niestety, automatyczne rozpoznanie tekstów wielu publikacji jest bardzo utrudnione. Wpływa na to kilka czynników: niska jakość materiału wejściowego spowodowana nieprecyzyjnym drukiem (zamazane znaki, duplikowane litery, niewielki rozmiar czcionki) czy papier o niskiej gramaturze, co powoduje efekt przebijania strony odwrotnej podczas skanowania. Złe warunki przechowywania (wilgotność) powodują marszczenie papieru, falowanie jego powierzchni, co z kolei sprawia, że tekst nie jest skanowany w linii prostej. Jednym z podstawowych sposobów wspomagających proces rozpoznania tekstu jest korekta słownikowa (często w czasie rzeczywistym). Biblioteki coraz częściej korzystają ze społecznej, zbiorowej korekty danych post-OCR (collaborative OCR text correction). Cechuje się ona m.in. wysoką trafnością oraz niskim nakładem kosztów.

OCR – opis technologii

Optyczne rozpoznawanie znaków to technologia, która pozwala na konwersję dokumentów z postaci obrazu rastrowego do przeszukiwalnego i edytowalnego tekstu. Proces ten odbywa się przy użyciu oprogramowania OCR, które przypisuje wizualizacji graficznej określonego symbolu jego kod z tablicy znaków – w przypadku tekstów polskich są to najczęściej standardy ISO-8859-2 lub UTF-8. Wynik może być zapisany w pliku tekstowym lub tylko wyświetlony na ekranie komputera. Współczesne programy OCR potrafią również zachować więcej informacji o OCR-owanym dokumencie – krój czcionki, rozmiary marginesów oraz układ wszystkich elementów na stronie. Wynikowy tekst jest zapisywany wówczas w formacie DOC lub PDF.

Współczynnik prawidłowego rozpoznania tekstu jest uzależniony m.in. od jakości materiału wejściowego. Ważnym parametrem jest rozdzielczość – do większości zastosowań wystarcza 300 dpi. W wybranych przypadkach, kiedy skanujemy obraz zawierający detale lub tekst drukowany małym rozmiarem czcionki (6 punktów typograficznych lub mniej), wymagana jest wyższa rozdzielczość.

Częstokroć OCR jest procesem bardzo kosztownym, wiele instytucji jednak podejmuje się tego zadania, ponieważ dane po OCR-owaniu:

– dają możliwość generowania indeksów, dzięki czemu pełne teksty publikacji mogą stać się przeszukiwalne,

– są znacznie mniejsze w porównaniu do wejściowych danych rastrowych. Do przechowywania dokumentów będziemy potrzebowali mniej miejsca na pamięciach masowych. Również ich przesyłanie będzie trwało znacznie krócej,

– stają się edytowalne – możemy zmieniać ich zawartość, dopisywać, usuwać treści.

OCR publikacji o niskiej jakości

Większość zasobów polskich bibliotek cyfrowych stanowią materiały historyczne. Wynika to głównie z chęci zachowania i udostępnienia dziedzictwa narodowego. Pragniemy ocalić dla potomności zbiory, które często stopniowo ulegają degradacji. Wiele zasobów nie jest również udostępnianych czytelnikom ze względu na możliwość wyrządzenia dodatkowych szkód zbiorom. Zamiana na postać cyfrową i ich udostępnienie w Internecie daje możliwość dostępu znaczącej rzeszy czytelników. Liczącym się powodem są z pewnością również prawa autorskie i wydawnicze, które w stosunku do publikacji historycznych wygasły. Dużo łatwiej, bez dodatkowych formalności i związanych z tym kosztów, publikacje historyczne zamienić na postać cyfrową i udostępnić czytelnikom.

Stan publikacji historycznych jest niestety bardzo niski, co wywołuje często niezadowalający współczynnik rozpoznania OCR. Wśród głównych czynników wpływających na taki stan rzeczy należy wymienić:

• proces zwany zakwaszaniem papieru, którego powodem są dwa „udoskonalenia" wprowadzone w XIX w. – nowy sposób zaklejania papieru oraz zmiana surowca, z którego papier był wytwarzany. Od 1996 r. nie produkuje się już kwaśnego papieru, pozostaje jednak problem wielu publikacji historycznych, które ulegają powolnej degradacji. Z jednej strony, wydawnictwa rozpadają się w rękach, z drugiej – należy je utrwalić dla potomnych. O skali problemu świadczy chociażby fakt, że w Bibliotece Jagiellońskiej 82% druków zwartych, czyli ok. 1 500 000 pozycji, zostało wydrukowanych na kwaśnym papierze (Barański 2006),

• niska gramatura papieru używanego w druku gazet (poniżej 80 g/m^2) powoduje podczas skanowania powstawanie zjawiska przebijania zawartości strony odwrotnej (Bednarek 2008),

• nieprecyzyjny druk – wiele gazet, szczególnie regionalnych, wydawanych w małych miejscowościach, było drukowanych na wysłużonych maszynach drukarskich. Już wkrótce po wydrukowaniu treść była mało czytelna. Dodatkowe szkody poczyniła degradacja powodująca powstawanie przerwań (nieciągłości) w literach. Do występujących często błędów drukarskich zaliczyć można również podwójne odbicie tekstu, niedodrukowane miejsca i zafarbowania,

• niewłaściwe przechowywanie. Jak zauważa Grzegorz Bednarek (Bednarek 2008): „Zwyczajowo, roczniki gazet oprawione w sztywną oprawę przechowywane są tak, że grzbiet oprawy widoczny jest dla oczu bibliotekarza. Jest to normalne z punktu widzenia konieczności stałego odszukiwania w bibliotece określonego woluminu. Zaś z punktu widzenia nie obniżania jakości przechowywanych gazet, grzbiet musiałby być u góry, by nie dopuścić do trwałych odkształceń arkuszy. Aby ocenić jak w danej bibliotece przechowywany jest zasób czasopism, wystarczy ogląd części strony przeciwległej do zszytego boku. (...) Niestety, testy wykazały, że takie odkształcenia znacząco obniżają jakość rozpoznania tekstu OCR.",

• pofalowana powierzchnia gazet i książek. Powodem może być wysoka kwasowość papieru lub zmiany wilgotności pomieszczenia, gdzie przechowywane były publikacje. Pofalowanie stanowi poważny problem dla oprogramowania OCR,

• czcionki nieużywane obecnie. Do niedawna poważnym problemem było rozpoznawanie tekstów drukowanych historycznymi czcionkami (np. Gotykiem). Obecnie systemy OCR coraz lepiej radzą sobie z tym utrudnieniem,

• zniszczenia powstałe podczas użytkowania, m.in. poplamienia publikacji, ich przedarcie lub wytarcie.

Niska jakość publikacji historycznych wpływa na niezadowalający stopień ich rozpoznawania przez aplikacje OCR. Świadczą o tym badania przeprowadzone w British Library na dwóch historycznych bazach danych gromadzących gazety pochodzące z XIX oraz XVII i XVIII w. (Tanner 2009). Badaniami objęto ok. 1% z 2 mln stron zasobów. Wybierano fragmenty najwyraźniejsze, z ponadprzeciętną jakością. Wyniki pokazały bardzo niską rozpoznawalność tekstu na poziomie słów znaczących, a więc najbardziej cennych jednostek leksykalnych podczas indeksowania. W bazie The 19th Century Newspaper Project było to 68,4%, natomiast w przypadku starszych czaso-

pism, pochodzących z bazy Burney, trafność na poziomie słowa znaczącego wyniosła zaledwie 48,4%, co oznacza, że ponad połowa słów znaczących została rozpoznana niepoprawnie.

Wspomaganie procesu rozpoznawania tekstu

Aby zwiększyć stopień rozpoznania tekstu stosuje się kilka metod, które podzielić możemy na dwie podstawowe grupy: programowe oraz korektę.

Metody programowe są stosowane bezpośrednio po skanowaniu, a przed procesem rozpoznawania tekstu. Mają one za zadanie polepszenie jakości materiału rastrowego, a tym samym zwiększenie współczynnika rozpoznania tekstu. Do podstawowych metod programowych należą:

Poziomowanie (ang. de-skew)

Metoda ta jest stosowana do materiałów, które zostały zeskanowane ukośnie. Powodem może być nieprecyzyjne ułożenie materiału w skanerze, ale tego typu błąd może również powstać podczas druku. Procedura ta musi być wykonana przed OCR-em – w przeciwnym wypadku współczynnik rozpoznania tekstu będzie obniżony.

Usuwanie szumów (ang. noise-removal)

Podczas skanowania mogą powstawać różnego rodzaju „szumy", wywołane m.in. przez kurz, brud lub inne zanieczyszczenia znajdujące się na szybie skanera lub obiektywie kamery cyfrowej. Wiele zanieczyszczeń może się znajdować na samym dokumencie – uwaga ta dotyczy szczególnie dokumentów historycznych. Szumy można oczywiście usuwać ręcznie, jest to jednak proces bardzo czaso- i kosztochłonny. W automatyczną eliminację zanieczyszczeń są wyposażone programy graficzne (np. Adobe Photoshop). Istnieje również grupa algorytmów odszumiających, które wykorzystuje oprogramowanie OCR.

Analiza układu strony (dokumentu) (ang. document layout analysis)

Zapis rastrowy nie przechowuje informacji o strukturze dokumentu, dlatego przed wykonaniem OCR-u konieczne jest dokonanie analizy układu dokumentu. Etap ten zawiera automatyczną identyfikację i klasyfikację elementów występujących na stronie.

Binaryzacja (ang. binarization)

Etap ten polega na zamianie obrazu rastrowego z odcieni szarości na obraz binarny (wartość piksela zapisywana jest na 1 bicie).

Korekta post-OCR

Korekta post-OCR bazuje na słownikach. Część systemów OCR (ABBYY Fine -Reader, ExperVision TypeReader & OpenRTK, OmniPage) jest wyposażona w słow-

niki w różnych językach. Do innych musimy dołączyć słownik zewnętrzny. Jakość korekty jest determinowana przez jakość słownika. Systemy OCR porównują każdy wyraz z rozpoznawanego tekstu z wystąpieniami w słowniku. Może się zdarzyć, że wyraz zostanie błędnie rozpoznany, a jego wystąpienie zostanie znalezione w słowniku. Zjawisko to określane jest mianem „false friend". Idealny (czysto hipotetycznie) słownik to taki, który zawiera wyłącznie wyrazy znajdujące się w sprawdzanym tekście. Jeżeli będzie liczył ich mniej, wówczas część wyrazów nie zostanie rozpoznana. Jeśli zaś słownik będzie zbyt obszerny – sprawdzanie wydłuży się. Dodatkowo zbyt duża liczba podpowiedzi może spowodować, że zostanie wybrana nieodpowiednia propozycja.

Korekta post-OCR może odbywać się w dwóch trybach (Hauser 2007):

1. **Półautomatyczna** – tekst jest rozpoznawany i porównywany ze słownikiem, a następnie wyrazy nierozpoznane zostają wyróżniane. W kolejnym etapie dla wyrazów nierozpoznanych są wyświetlane propozycje. Konieczna jest ingerencja operatora (człowieka), który wybiera odpowiednie wyrażenie.

2. **Automatyczna** – w pełni automatyczna korekta nie wymaga obecności człowieka. Zaletą takiego rozwiązania jest szybkość działania – duże zbiory danych mogą zostać poprawione w krótkim czasie. Automatyczna korekta post-OCR odbywa się w trzech etapach: tokenizacja – system OCR generuje wyrazy, wyszukiwanie słów w słowniku, korekta – zostają podstawione odpowiednie terminy ze słownika.

Crowdsourcing

Coraz częściej do pracy w sieci wykorzystuje się społeczności. Jeff Howe użył w czerwcu 2006 r. w artykule do magazynu „Wired" neologizmu „crowdsourcing". Oznacza on przypisywanie tradycyjnych obowiązków pracowników najemnych grupie (*crowd* – z ang. tłum; *sourcing* – z ang. czerpanie źródeł) ludzi czy społeczności (Crowdsourcing, 2011).

Podstawowe korzyści płynące z crowdsourcingu:

– problemy mogą być rozwiązane przy względnie niskich kosztach, na ogół bardzo szybko,
– płaci się w zależności od rezultatu, czasem wręcz nagradzanie się pomija,
– organizacja dociera do szerszego grona talentów niż tylko członkowie samej organizacji,
– poprzez słuchanie społeczności organizacja zyskuje informacje z pierwszej ręki o potrzebach i pragnieniach klientów,
– społeczność może poczuć, że współtworzy markę.

Crowdsourcing wykorzystują w swoich działaniach coraz częściej firmy komercyjne i instytucje, w tym biblioteki – również do prowadzenia korekty post-OCR.

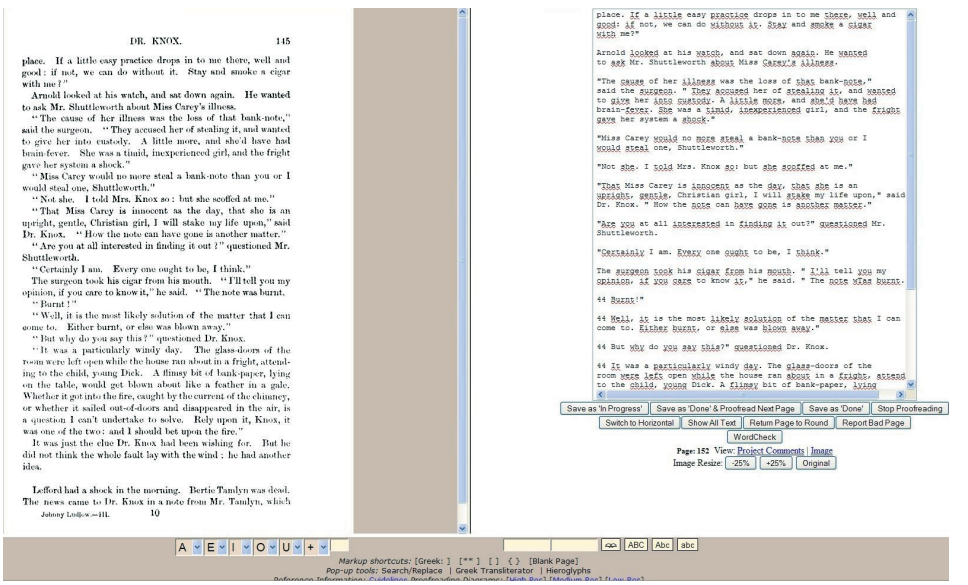
Korekta społeczna

Distributed Proofreaders

Do pionierów, którzy w swoich działaniach wykorzystywali społeczności, należą twórcy projektu Distributed Proofreaders (http://www.pgdp.net/c/) (Distributed Proofreaders, 2011). Celem rozpoczętego przez Charlesa Franksa w 2000 r. projektu było wsparcie Projektu Gutenberg. Publikacje o statusie „Public Domain" są najpierw skanowane, a następnie OCR-owane. Ich jakość jest często niezadowalająca, dlatego teksty przed publikacją muszą być korygowane. Dokonują tego wolontariusze, pracując nad książkami w sieci. W przeglądarce internetowej obok zeskanowanej strony pojawia się tekst do korekty. Projekt Distributed Proofreaders posiada własne oprogramowanie, a także forum i wiki. Każda strona jest sprawdzana wielokrotnie, a następnie udostępniana w Internecie.

Nad jedną książką pracuje jednocześnie wielu korektorów. W projekcie mogą uczestniczyć osoby, które się zarejestrują. W procesie rejestracji należy podać swoje imię, nazwisko, wybrać nazwę użytkownika i hasło oraz podać adres e-mail. Po weryfikacji adresu e-mail można przystąpić do partycypacji w projekcie. Distributed Proofreaders wyróżnia trzy fazy korekty publikacji:

Pierwsza faza, **brązowa**, obejmuje korektę publikacji, które trafiły do systemu po optycznym rozpoznawaniu znaków. Każdy z tekstów jest trzykrotnie korygowany (Proofreading Round 1-3), a następnie trafia do etapu formatowania. Po jego ukończeniu publikacje przekazywane są do **drugiej fazy**, **srebrnej** (ang. Completed Silver



Ryc. 1. Projekt Distributed Proofreaders: strona publikacji w dwóch wersjach: rastrowej (skanu strony) oraz tekstu w postaci znakowej do poprawki

Title	Author	Language	Genre	Project Manager	Available Pages	Total Pages	Days
*The trangression of Andrew Vane [Part 1 of 3]	Carryl, Guy Wetmore	English	BEGINNERS ONLY General Fiction	BEGIN	116	116	0
*There is no death [Part 3 of 3]	Marryat, Florence	English	BEGINNERS ONLY Non-Fiction	BEGIN	5	100	2
Gli amori -- 15 -- Anacronismo	Federico De Roberto	Italian	BEGINNERS ONLY Short Story	BEGIN	3	8	3
*Histoire de ma Vie 85/87	George Sand	French	BEGINNERS ONLY Biography	BEGIN	0	13	5
*Geschichten [1914] {Fraktur} (Teil 10 von 11)	Walser, Robert	German	BEGINNERS ONLY Short Story	BEGIN	5	21	9
The Maker of Opportunities	Gibbs, George	English	Romance	DACSoft	293	293	0
Little Man's family: pre-primer	Enochs, J. B.	English	EASY Juvenile	JulietS	34	38	0
Every Boy's Book: A Complete Encyclopædia of Sports and Amusements.	Routledge, Edmund	English	Juvenile	Mebyon	849	864	0
History of the reformation in the sixteenth century (volume 4)	Merle d'Aubignâe, J. H.	English	History	crb11	427	507	1
Spectra	Emanuel Morgan and Anne Knish	English	EASY Poetry	dvdeug	0	78	1
REVISIONE RAPIDA-Passeggiate per l'Italia vol. 4--Palermo (2)	Ferdinand Gregorovius	Italian	Travel	garweyne	33	40	1
The Diatomaceæ of Philadelphia and vicinity {P1->P1}	Boyer, Charles Sumner	English	Biology	puppernutter	207	226	1
The Archaeology of the Yakima Valley	Smith, Harlan Ingersoll	English	Archaeology	tunelera	163	208	1
The trail of the axe	Culhan, Ridgwell	English	General Fiction	bunny-crunch	174	428	1
Johnny Ludlow; Series 3	Wood, Henry	English	Mystery	De2164	330	480	1
Storia delle repubbliche italiane dei secoli di mezzo, v. 11/16	Simondo Sismondi	Italian	History	paganelli	442	445	2
Cyclopedia of Commerce, Accountancy, Business Administration, v. 5	Corps of Professionals	English	Business	JulietS	341	366	2
With the Battle Fleet	Matthews, Franklin	English	Military	Quister	269	352	2
The Esperantist, Volume 2, No. 3 (March 1905) {P1->P1}	Mudie, H. Bolingbroke (editor)	Esperanto with English	Periodical	dvdeug	40	65	3
Modern Leaders: Being a Series of Biographical Sketches	McCarthy, Justin	English	Biography	donovan	177	243	3
Peggy Owen at Yorktown	Lucy Foster Madison	English	EASY Juvenile	JulietS	0	421	3
The last straw {P1->P1}	Titus, Harold	English	Western	bunny-crunch	102	292	3
Dal mio verziere	Jolanda	Italian	Other	garweyne	163	265	3
Il ferro	Gabriele D'Annunzio	Italian	EASY Drama	garweyne	177	209	3
Strandingshistorier	Mylius-Erichsen, Ludvig	Danish	Adventure	hanne	182	212	4
The Columbia River: Its History, Its Myths, Its Scenery, Its Commerce {P1->P1}	Lyman, William Denison	English	Geography	ms_e	64	514	4
Letters of Felix Mendelssohn Bartholdy from Italy and Switzerland. {P1->P1}	Mendelssohn-Bartholdy, Felix	English	Correspondence	hdmtrad	160	378	4
La plebe (4/4)	Vittorio Bersezio	Italian	General Fiction	garweyne	289	519	4

Ryc. 2. Projekt Distributed Proofreaders: lista publikacji przygotowanych do korekty

E-Texts). W trakcie tej fazy wolontariusz przeprowadza ostateczną kontrolę tekstu w celu sprawdzenia jego spójności i poprawności. **Ostatnia faza** to „ukończone złote teksty elektroniczne" (ang. Completed Gold E-Texts). Status ten otrzymują publikacje, które przeszły wszystkie fazy: korektę, formatowanie oraz przetwarzanie końcowe. W takiej postaci są przekazywane do Projektu Gutenberg i mogą zostać wyświetlone przez użytkowników w przeglądarce internetowej oraz załadowane do komputera.

Każdemu zalogowanemu użytkownikowi, po wyborze etapu, pojawia się na ekranie komputera lista dostępnych publikacji z podstawowymi informacjami bibliograficznymi (tytuł, imię i nazwisko autora, język publikacji), a także informacje o już skorygowanych lub sformatowanych stronach oraz ogólnej liczbie stron w publikacji.

Po kliknięciu na tytule pojawia się strona publikacji w dwóch wersjach: rastrowej (skanu strony) oraz tekstu w postaci znakowej do poprawki. Użytkownik może zmieniać układ obu elementów składowych, w zależności od tego, jakim monitorem dysponuje. Dla coraz popularniejszych monitorów o stosunku boków 16:9 wygodniejszy będzie układ poziomy, gdzie zeskanowana strona jest po lewej stronie, a tekst do poprawki po prawej.

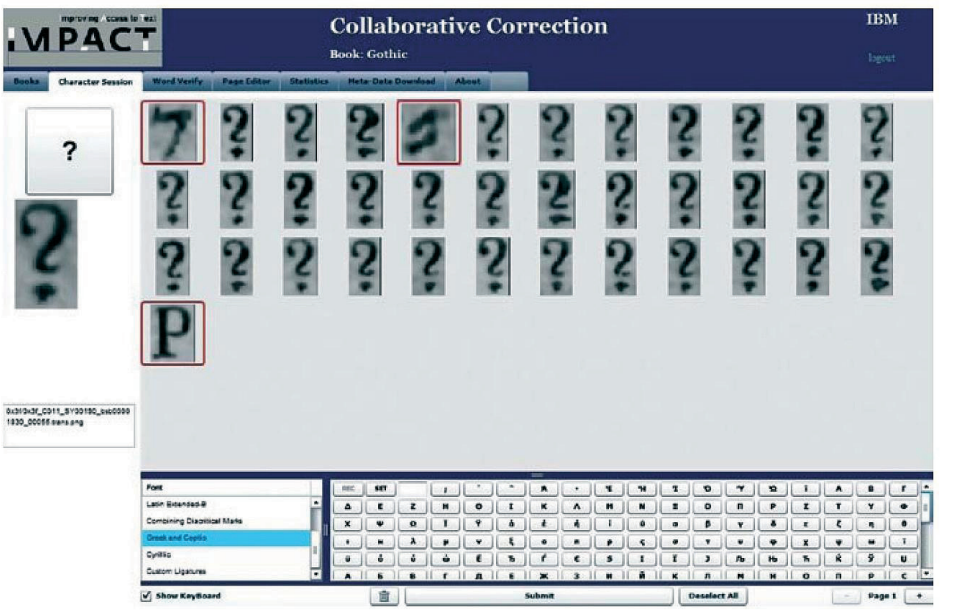
Korektor może korzystać z dodatkowych narzędzi. Dostępna jest lista rozwijalna, z której wybiera się znaki diakrytyczne niedostępne na standardowej klawiaturze ze znakami alfabetu łacińskiego. Dodatkowym ułatwieniem jest możliwość korzystania ze słowników wspomagających korektę. Są one dostępne nie tylko w języku angielskim, ale również w 21 innych językach. Korektor można również powiększać lub pomniejszać rozmiar zeskanowanej strony. Dostępnych jest wiele dokumentów po-

mocniczych, które wspomagają proces korygowania z podstawową regułą: „Don't change what the author wrote!" Zawierają one liczne przykłady poprawnie skorygowanych stron.

Do końca stycznia 2011 r. w systemie zarejestrowało się ponad 101 tys. użytkowników, którzy skorygowali 19541 publikacji. Warto wspomnieć, że obok ebooków w języku angielskim korygowane są teksty w kilkudziesięciu innych językach, w tym również w języku polskim.

Projekt IMPACT

Instytucje europejskie również zauważyły problem optycznego rozpoznawania tekstów. Jedną z inicjatyw finansowanych przez Komisję Europejską jest projekt IMPACT, który położył nacisk na zwiększenie efektywności procesu korekty poprzez udział wolontariuszy. W jego ramach utworzono platformę bazującą na przeglądarkach internetowych o nazwie CONCERT – COllaborative eNgine for Correction of ExtRacted Text (Neudecker, Tzadok 2010). System posiada własne narzędzia OCR. Po zeskanowaniu dokumentów oraz rozpoznaniu tekstu zostaje on zaprezentowany w postaci tzw. carpets session. Platforma CONCERT wyróżnia trzy poziomy korekty. Jako pierwsze na ekranie zostają wyświetlone wszystkie wystąpienia określonego znaku. Są to litery, które zostały przez system oznaczone jako „podejrzane", a więc takie, które mogły zostać niepoprawnie rozpoznane. W oknie przeglądarki internetowej użytkownik może zobaczyć wszystkie wystąpienia określonego znaku i zaznaczyć te, które zostały błędnie mu przypisane. Tym samym zaznaczone przez operatora znaki



Ryc. 3. Platforma CONCERT: operator zaznacza błędnie rozpoznane znaki

zostają odrzucone, pozostałe natomiast zostaną zaaprobowane i przyjęte do publikacji jako poprawne. Działania operatora są również traktowane jako wskazówki dla adaptacyjnego silnika OCR dla określonego fontu[1].

Może się jednak zdarzyć, że znak jest bardzo nieczytelny i operatorowi trudno jest rozstrzygnąć, z jakim znakiem ma do czynienia. Wówczas przenosi się na kolejny poziom: 'word session', na którym może wyświetlić wystąpienie znaku w określonym kontekście (wyrazie). Po rozpoznaniu wyraz jest dodawany do słownika, który ułatwia korektę dokumentu.

Trzeci, najwyższy poziom to pełna strona. Jeżeli cały wyraz jest nieczytelny, wówczas operator może wyświetlić pełną, zeskanowaną stronę. Ten poziom jest szczególnie przydatny w przypadku błędnie połączonych lub rozdzielonych jednostek leksykalnych.

System ma również wykrywać „żartownisiów", którzy podczas korekty celowo wprowadzają błędy do tekstów, oraz motywować najbardziej zaangażowanych w projekt wolontariuszy. Zaplanowano również wprowadzanie celowych błędów do systemu, aby zidentyfikować wolontariuszy, którzy sumiennie podchodzą do pracy, oraz tych, którzy pozostawiają w tekście wiele błędów.

reCAPTCHA

Na szczególną metodę korekty tekstów post-OCR wpadli twórcy projektu reCAPTCHA (Ahn 2008, Science). Postanowili zmodyfikować technologię stosowaną do weryfikacji osób i maszyn – CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart). CAPTCHA stosowana jest często podczas przesyłania danych za pomocą formularzy. Test ten chroni przed masowym przesyłaniem wiadomości przez automaty. Spotykamy go przy zakładaniu kont w serwisach, forach dyskusyjnych, wysyłaniu wiadomości czy wpisywaniu komentarzy. CAPTCHA stosowana jest najczęściej w postaci wyświetlanych znaków (cyfr, liter) lub wyrazów w postaci rastrowej. Jest ona poddana zniekształceniom, co ma chronić serwis internetowy przed użyciem oprogramowania OCR. System ten jest używany ponad 100 milionów razy dziennie.

ReCAPTCHA spełnia identyczną funkcję co wspomniana wyżej CAPTCHA, dodatkowo jednak pomaga w korekcie OCR-owanego tekstu. Zamiast jednego proszeni jesteśmy o przepisanie dwóch wyrazów. Jeden z nich jest wyrazem kontrolnym, służącym weryfikacji, drugi natomiast – wyrazem pochodzącym z korekty post-OCR.

Twórcy projektu do rozpoznania tekstu na etapie OCR-u wykorzystywali dwa niezależne programy OCR. Każdy wyraz, który został rozpoznany różnie przez każdy z dwóch programów, albo taki, którego nie ma w słowniku języka angielskiego, został oznaczony jako „podejrzany". Wedle statystyk 96% wyrazów podejrzanych jest rozpoznawana nieprawidłowo przynajmniej przez jeden z programów OCR. Podejrzany wyraz jest zapisywany jako obraz i podsyłany z drugim wyrazem, którego znaczenie

[1] Neudecker, C., Tzadok A., User Collaboration for Improving Access to Historical Texts, Liber Quarterly 20(1), September 2010.

Ryc. 4. Projekt reCAPTCHA – użytkownik przepisuje dwa wyrazy: wyraz kontrolny służący weryfikacji oraz wyraz pochodzący z korekty post-OCR

jest znane. Użytkownik jest proszony o wpisanie dwóch wyrazów: kontrolnego oraz wyrazu nieznanego. Ten sam wyraz nieznany jest wysyłany do różnych użytkowników, za każdym razem z innym zniekształceniem. Jeżeli trzy pierwsze osoby rozpoznają wyraz tak samo, staje się on wyrazem kontrolnym. W przeciwnym wypadku (jeżeli zaistnieją rozbieżności) system przesyła wyraz nieznany kolejnym użytkownikom. Jeżeli sześciu użytkowników zażąda nowej pary wyrazów, wówczas wyraz zostaje oznaczony jako „nieczytelny".

System reCAPTCHA okazał się być bardzo skuteczny. Do celów statystycznych wybrano 50 przypadkowych artykułów z czasopisma „New York Times" z różnych lat (1860, 1865, 1908, 1935 i 1970). Ogólna liczba słów do rozpoznania wyniosła 24080. Po podliczeniu wyników okazało się, że system działa ze skutecznością 99,1% (216 błędów na 24080 słów). Skuteczność systemu OCR (bez korekty) wyniosła 83,5% (3976 błędów). Po roku funkcjonowania systemu użytkownicy rozpoznali 1,2 miliarda CAPTCHA, pomagając rozszyfrować ponad 444 milionów podejrzanych wyrazów, co daje średnio 17600 książek.

Australian Newspapers Digitisation Program

Rose Holley w artykule „Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers" (Holley 2009) opisuje społeczną korektę tekstu post-OCR wykonywaną w ramach projektu ANDP (Australian Newspapers Digitisation Program). Celem projektu jest digitalizacja australijskich czasopism historycznych datowanych na lata 1803–1954. Nie są one obciążone prawami autorskimi, dzięki czemu można udostępnić czytelnikom pełne teksty. Uczestnicy projektu od początku borykali się z niską jakością materiałów wyjściowych. Pierwsze australijskie gazety były drukowane na maszynach drukarskich, które zostały wycofane z użycia w Anglii. Problemem był również brak odpowiedniego papieru. Wyniki projektu zostały udostępnione w wersji beta w lipcu 2008 r. Zasób liczył wówczas 3,5 miliona artykułów (360 tys. stron czasopism). Docelowo w 2011 r. ma być przygotowanych 40 mln artykułów (4,4 mln stron czasopism). Dzięki zastosowa-

niu korekty społecznej udało się poprawić część tekstów. Początkowo grupa społecznych korektorów liczyła 1300 osób. W ciągu pierwszych 6 miesięcy istnienia systemu skorygowali oni 2 miliony wierszy tekstu w 100 000 artykułach. Skorygowane dane opublikowano w styczniu 2009 r. Holley zwraca uwagę na istotnej zalety, ale również zagrożenia związane ze społeczną korektą tekstów:

– potencjalny wandalizm tekstu (korektorzy będą celowo wprowadzali błędy do tekstu),

– duże zużycie bazy/zasobów serwerów z powodu masowego dostępu,

– użytkownicy nie zechcą korygować tekstów i czas przeznaczony na korektę będzie zmarnowany,

– użytkownicy nie zrozumieją koncepcji korygowania tekstu post-OCR,

– użytkownicy „wystraszą się" pełnego błędów, niekiedy mało zrozumiałego tekstu.

Zakończenie

Wiele bibliotek boryka się z problemem poprawnego, optycznego rozpoznania publikacji. Jedną z najbardziej obiecujących metod jest korekta społeczna. Zapewnia bardzo wysoką skuteczność, może być tania czy wręcz darmowa. Czytelnicy są nie tylko biernymi użytkownikami zbiorów, ale czynnie angażują się w ich powstawanie. Działania te z pewnością przyczyniają się do szybszego powstawania pełnowartościowych bibliotek cyfrowych. Również w Polsce warto rozpropagować tę ideę.

BIBLIOGRAFIA

Ahn von L., Maurer B., McMillen B., Abraham D., Blum M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science vol. 321, s. 1465–1468.

Barański Andrzej (2006). Kwaśny papier, http://www3.uj.edu.pl/alma/alma/82/09.pdf [odczyt: 30.01.2011].

Bednarek Grzegorz (2008). Format DjVu a problem digitalizacji gazet i czasopism, http://www.djvu.com.pl/galeria/UJ/Gazety_czasopisma.php [odczyt: 30.01.2011].

Crowdsourcing. [In:] Wikipedia. The free encyclopedia, http://pl.wikipedia.org/wiki/Crowdsourcing [odczyt: 30.01.2011].

Distributed Proofreaders. [In:] Wikipedia. The free encyclopedia, http://en.wikipedia.org/wiki/Distributed_Proofreaders [odczyt: 30.01.2011].

EDL Report on Digitisation in European National Libraries 2006-2012 (2008), http://www.cenl.org/docs/Report_digitisation_NLs.pdf [odczyt: 30.01.2011].

Hauser W. Andreas (2007). OCR Postcorrection of Historical Texts, http://www.cip.ifi.lmu.de/~hauser/papers/histOCRNachkorrektur.pdf [odczyt: 30.01.2011].

Holley Rose (2009). Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers, http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf [odczyt: 30.01.2011].

Impact: Improving access to text: Concept, http://www.impact-project.eu/about-the-project/concept/ [odczyt: 30.01.2011].

Lin Leo (2009). Improving Digital Library Support for Historic Newspaper Collections, http://research-commons.waikato.ac.nz/bitstream/10289/3262/1/thesis.pdf [odczyt: 30.01.2011].

Łotewski Tomasz. Kwaśny papier – chemiczna katastrofa w bibliotekach, http://kangur.uek.krakow.pl/biblioteka/konferencja/Aktualnosci/005.pdf [odczyt: 30.01.2011].

Neudecker C., Tzadok A. (2010). User Collaboration for Improving Access to Historical Texts. Liber Quarterly 20(1), September 2010.

Tanner S., Muñoz T., Ros P. H. (2009). Measuring Mass Text Digitization Quality and Usefulness, D-Lib Magazine, http://www.dlib.org/dlib/july09/munoz/07munoz.html [odczyt: 30.01.2011].

Vamvakas G., Gatos B., Stamatopoulos N., Perantonis S.J. (2008). A Complete Optical Character Recognition Methodology for Historical Documents, The Eighth IAPR International Workshop on Document Analysis Systems, http://iit.demokritos.gr/~bgat/3337a525.pdf [odczyt: 30.01.2011].