Kwartalnik Językoznawczy 2015/3-4
DOI: 10.14746/kj.2015.3-4.1

Janusz S. Bień

Elektroniczny indeks do słownika Lindego[1]

1. Geneza indeksu

Podstawowym celem indeksu jest ułatwienie korzystania z dygitalizacji słownika Lindego opracowywanej w Katedrze Lingwistyki Formalnej Uniwersytetu Warszawskiego z inicjatywy autora i pod jego kierunkiem. Jego potencjalne zastosowania są jednak znacznie szersze, dzięki czemu część prac nad indeksem mogła zostać sfinansowana przez projekt IMPACT[2]. Punktem wyjścia był opublikowany w 1965 roku „Indeks *a tergo* do Słownika języka polskiego S.B. Lindego”[3].

2. Indeks *a tergo* do słownika S.B. Lindego

2.1. Historia

Był to pierwszy indeks *a tergo* dla języka polskiego, stąd najwyraźniej odczuwano potrzebę uzasadnienia celowości tego przedsięwzięcia przez przywołanie poglądów Onufrego Kopczyńskiego, Jana Baudouina de Courtenaya oraz już opublikowanych indeksów *a tergo* dla innych języków słowiańskich.

[1] Omawiane prace były częściowo finansowane przez unijny projekt IMPACT (*IMProving ACcess to Texts*, FP7-ICT7 215064).

[2] J.S. Bień, *Polskie zasoby językowe w projekcie IMPACT*, 2011; https://www.slideshare.net/jsbien/polskie-zasoby-jzykowe-w-projekcie-impact [dostęp: 24.05.2018].

[3] R. Grzegorczykowa i in., *Indeks a tergo do Słownika języka polskiego S.B. Lindego*, red. W. Doroszewski, 1965; http://ebuw.uw.edu.pl/publication/339849 [dostęp: 9.03.2018].

Profesor Renata Grzegorczykowa dostarczyła mi następujących informacji na temat genezy indeksu i jego opracowania (mejl z 28 kwietnia 2014 roku):

> Pracownia ta[4] miała za zadanie opracowanie pod kątem słowotwórczym zasobów słownikowych, zebranych dla potrzeb powstającego wielkiego słownika (SJPD). Sporządzenie indeksów *a tergo* (do słownika Lindego – 1965, a następnie do SJPD – 1973) dawało wstępną bazę materiałową dla opisu słowotwórczego (rozumianego genetycznie), a także dodatkowo dla obserwacji fleksyjnych. Próbkę takiego słowotwórczego opisu przedstawiał *Zeszyt próbny indeksu słowotwórczego do „Słownika języka polskiego"* pod redakcją Witolda Doroszewskiego, 1963.
>
> W opracowaniu indeksu *a tergo* do Lindego prof. Doroszewski praktycznie nie brał żadnego udziału, poza tym, że kierował Pracownią. Indeks opracował trzyosobowy zespół Pracowni: R. Grzegorczykowa, Zofia Kawyn-Kurzowa i Jadwiga Puzynina, która *de facto* była osobą kierującą pracami redakcyjnymi. Pracę benedyktyńską rozpisania haseł na kartkach i ułożenia *a tergo* wykonały osoby w ramach tzw. prac zleconych. Po wykonaniu pracy kartki zostały zniszczone, nie było bowiem na nich żadnych interesujących informacji.

Indeks został wydany przez Wydawnictwa Uniwersytetu Warszawskiego, a z obowiązkowej w tym okresie metryki książki można się dowiedzieć, że nakład wynosił 500 egzemplarzy (plus 25 gratisów, jeśli dobrze interpretuję zapis), a cena 62 zł.

2.2. Status prawny

Na stronie tytułowej czytamy: „Uniwersytet Warszawski", następnie „INDEKS *A TERGO* do SŁOWNIKA JĘZYKA POLSKIEGO S. B. LINDEGO pod redakcją Witolda Doroszewskiego". Na dole znajdują się godło uczelni, nazwa wydawnictwa i rok wydania. Na odwrocie strony tytułowej znajduje się napis: „Opracowały: R. Grzegorczykowa, Z. Kurzowa, J. Puzynina". Pod wstępem (s. 7) znajduje się podpis: „Pracownia Leksykologiczna przy Katedrze Języka Polskiego UW" (na doklejonym pasku papieru – widocznie został omyłkowo pominięty w druku).

Podstawowe pytanie dotyczy tego, czy publikacja ta stanowi utwór w sensie prawa autorskiego. Pierwsza polska ustawa o prawie autorskim z 29 marca 1926 roku stwierdzała w art. 1:

---

[4] Pracownia Leksykologiczna przy Katedrze Języka Polskiego Uniwersytetu Warszawskiego – J.S.B.

Przedmiotem prawa autorskiego jest od chwili ustalenia w jakiejbądź postaci (słowem żywem, pismem, drukiem, rysunkiem, barwą, bryłą, dźwiękiem, mimiką, rytmiką) każdy przejaw działalności duchowej, noszący cechę osobistej twórczości.

Współcześnie (w ustawie z 4 lutego 1994 roku z późniejszymi zmianami) artykuł ten brzmi:

Przedmiotem prawa autorskiego jest każdy przejaw działalności twórczej o indywidualnym charakterze, ustalony w jakiejkolwiek postaci, niezależnie od wartości, przeznaczenia i sposobu wyrażania (utwór).

Czy indeks stanowi przejaw działalności twórczej o indywidualnym charakterze? Nie można tego wykluczyć, jak pokazuje spór między Wydawnictwem C.H. Beck i Oficyną Wydawniczą Verba na temat praw autorskich do specyficznych wyróżnień typograficznych, zakończony w 2006 roku ugodą sądową (syg. IX GC 400/05)[5].

Zakładamy zatem, że jest to utwór – kto w takim razie jest jego autorem? Zgodnie z art. 8 pkt 2 prawa autorskiego domniemywa się, że „twórcą jest osoba, której nazwisko w tym charakterze uwidoczniono na egzemplarzach utworu lub której autorstwo podano do publicznej wiadomości w jakikolwiek inny sposób w związku z rozpowszechnianiem utworu". Moim zdaniem można zatem przyjąć, że współautorami są Witold Doroszewski (zmarł w 1976 roku), Renata Grzegorczykowa, Zofia Kurzowa (początkowo podpisywała się Kawyn-Kurzowa; zmarła w 2003 roku) i Jadwiga Puzynina. Traktowanie utworu jako zbiorowego w sensie art. 11, do którego prawa przysługują producentowi lub wydawcy, nie byłoby w tym przypadku właściwe. Pojęcie utworu pracowniczego zostało wprowadzone do polskiego prawa autorskiego znacznie później i w związku z tym też nie ma tutaj zastosowania.

Nie wiadomo, czy została sporządzona umowa wydawnicza na opublikowanie indeksu. Gdyby jednak taka umowa istniała, to wiadomo, jaka byłaby jej treść – obowiązywał wówczas ogólnopolski wzorzec. Po wyczerpaniu nakładu autorzy mieli prawo wezwać wydawnictwo do sporządzenia kolejnego wydania, a w razie odmowy wszystkie prawa wracały do autorów; z drugiej strony wydawnictwo miało prawo do kolejnych wydań do czasu formalnego rozwiązania umowy.

---

[5] Por. R. Horbaczewski, *Nagłówki przepisów muszą być różne*, „Rzeczpospolita" 2006; http://archiwum.rp.pl/artykul/607574-Naglowki-przepisow-musza-byc-rozne.html [dostęp: 9.03.2018].

Istotną kwestią było więc poznanie stanowiska wydawnictwa – w piśmie z 10 września 2010 roku ówczesny dyrektor Wydawnictw Uniwersytetu Warszawskiego Ryszard Burek oświadczył, że Wydawnictwa nie roszczą sobie żadnych praw do tego utworu.

W sytuacji, kiedy autorzy (lub ich spadkobiercy) dysponują pełnią praw autorskich, powstaje pytanie, jaki z tych praw zrobić użytek. Dla rozwoju nauki najlepsze jest, jeśli udzielą oni licencji pozwalającej na wykorzystywanie ich dorobku w dalszych pracach. Ponieważ sformułowanie takiej licencji nie jest proste, popularne jest używanie różnych gotowych wzorów. W przypadku indeksu zaproponowałem użycie jednej z licencji *Creative Commons* (co moim zdaniem można tłumaczyć jako *Wspólnota Twórcza*), mianowicie wariantu nazywanego skrótowo *Uznanie autorstwa – Na tych samych warunkach*. Jego zasady można streścić następująco:

a) wolno:

– kopiować i rozpowszechniać utwór,

– tworzyć i rozpowszechniać utwory zależne (pochodne),

b) pod warunkiem:

– oznaczenia autorstwa,

– rozpowszechniania utworu oryginalnego i utworów zależnych tylko na zasadach takiej samej licencji.

Pełny tekst licencji jest dostępny na witrynie międzynarodowej organizacji Creative Commons (http://creativecommons.org/licenses/by-sa/3.0/pl/legalcode). Z właścicielami autorskich praw do indeksu kontaktowałem się sukcesywnie: pierwsze oświadczenia otrzymałem w lutym 2009 roku, a ostatnie trzy lata później.

2.3. Dygitalizacja

Pierwotnie indeks planował zdygitalizować Tadeusz Piotrowski, który wspominał o tym w prywatnej korespondencji jeszcze w styczniu 2006 roku, a w 2007 roku zlecił wykonanie skanów indeksu, które potem mi udostępnił. Skany niestety nie były zbyt dobrej jakości, więc w 2010 roku wskanowałem egzemplarz indeksu znajdujący się w Katedrze Lingwistyki Formalnej UW i – po dopełnieniu formalności, o których mowa wyżej – udostępniłem w bibliotece cyfrowej Katedry. Okazało się jednak, że ten skan również nie jest w pełni zadowalający i w związku z tym w razie potrzeby można rozważać wskanowanie indeksu po raz kolejny. Ze względu na defekty matrycy powtarzające się we wszystkich zapewne egzemplarzach i niską jakość papieru niektórych liter trzeba się niestety domyślać na

podstawie kontekstu nawet przy pracy z drukowanym oryginałem. Istnieje techniczna możliwość stworzenia wersji elektronicznej z odpowiednimi poprawkami i komentarzami, ale wydaje się, że nie ma zapotrzebowania na takie krytyczne wydanie indeksu.

Do optycznego rozpoznawania znaków wykorzystano popularny komercyjny program ABBY FineReader 10 Professional (tzw. wersja Desktop) – w tym czasie była to zdecydowanie najlepsza możliwość. Wyniki zapisywano w formacie PDF. Format ten zawiera m.in. informacje o wielkości fontów i układzie strony, ale nie są one bezpośrednio dostępne. W celu wykorzystania tych danych niezbędne było wykorzystanie narzędzi programistycznych stworzonych przez Tomasza Olejniczaka w ramach pracy magisterskiej na kierunku informatyka[6]. Okazało się jednak, że program FineReader w ogóle nie radził sobie z rozpoznaniem układu strony (być może z powodu nietypowego wyrównania łamów do prawej), a z rozpoznaniem wielkości czcionek (stopnia pisma) – odgrywającej bardzo istotną rolę w indeksie, bo odróżniającej hasła od podhaseł (por. pkt 3.4) – też miał sporo problemów.

W ramach projektu IMPACT Olejniczak napisał kilka prostych programów wykorzystujących zawartą w indeksie redundancję informacji do wykrywania niektórych typów błędnego rozpoznania znaków – wykorzystał do tego krótkie, ale bardzo pożyteczne niepublikowane opracowanie Joanny Bilińskiej *Opis Indeksu a tergo do Słownika języka polskiego S.B. Lindego* (10.03.2010, 4 s.). Najważniejszą funkcją programu było sygnalizowanie zakłócenia porządku alfabetycznego haseł, co mogło mieć trojakie przyczyny: błąd rozpoznania hasła, błędne potraktowanie podhasła jako hasła oraz – bardzo rzadkie – pomyłka redaktorów. Niestety nie ma praktycznie sposobu automatycznego odróżnienia dwóch pierwszych przypadków. Co więcej, w przypadku pojawiania się w niewłaściwej kolejności dwóch haseł nie ma prostego sposobu odróżnienia, które z nich jest błędne; choć teoretycznie jest to w pewnym stopniu możliwe, nie dysponowaliśmy czasem niezbędnym do stworzenia odpowiednio wyrafinowanego programu. W rezultacie program czasami interpretował dane niewłaściwie i po zakłóceniu porządku alfabetycznego przez błędne hasło lawinowo interpretował jako błędne następujące po nim hasła poprawne. Pomimo tych wad program istotnie ułatwił przeprowadzenie ręcznej

[6] T. Olejniczak, *Obsługa formatu PDF/A na potrzeby dygitalizacji tekstów*, niepublikowana praca magisterska, 2011, Wydział Matematyki i Informatyki Uniwersytetu Warszawskiego; https://bitbucket.org/jsbien/pdfautils-fork/downloads/mgr_to236111.pdf [dostęp: 24.05.2018].

korekty wykrytych błędów rozpoznania, którą wykonał niżej podpisany wspólnie z Moniką Kresą, zatrudnioną w Katedrze Lingwistyki Formalnej na czas realizacji projektu IMPACT.

W korektę indeksu wniósł wkład również Krzysztof Szafran, także zatrudniony w Katedrze Lingwistyki Formalnej na czas realizacji projektu IMPACT. Wykorzystał on w tym celu swój program analizy morfologicznej SAM[7] oraz analizator morfologiczny Morfeusz[8].

Oczywiście najlepszą formą korekty byłoby odszukanie haseł indeksu w słowniku. Pewne kroki w tym kierunku zostały zrobione w ramach projektu IMPACT z użyciem wstępnej dygitalizacji słownika, zawierającej niestety dużo błędów rozpoznania znaków. W konsekwencji za niewątpliwie poprawne można było uznać około 60 000 pozycji indeksu, czyli mniej więcej około 75% całości. Te częściowe wyniki zostały przekazane do projektu IMPACT i udostępnione w bibliotece cyfrowej KLF.

3. Pojęcia hasła w słowniku i indeksie *a tergo*

3.1. Hasła

W VI tomie słownika na stronach 24–37 znajduje się przedruk dokonanego przez Konstantego Wolskiego tłumaczenia niemieckojęzycznej recenzji pierwszego tomu – uzupełnionego o komentarze tłumacza – która ukazała się w 1808 roku w *Allgemeine Literatur-Zeitung* w Halle[9]. Na s. 31 przedruku czytamy (wyróżnienia moje):

> […] słowo wzięte do objaśnienia, **wersalikami** się różni, i jest na czele umieszczone; poczem następuje Polskie wyłuszczenie znaczenia, i Niemieckie tegoż słowa tłumaczenie, dalej wyrazy Czeskie, Słowackie, Windyjskie, Sorabskie, Rosyjskie, innych pobratymczych, nawet i obcych języków, które się dają z Polskim porównać. Po pierwiastkowem słowie, po klasyfikacji, po wyłuszczeniu i objaśnieniu wszystkich jego znaczeń, kładą się

[7] K. Szafran, *Analizator morfologiczny SAM-95: opis użytkowy*. TR 96-05 (226), Warszawa 1996, Instytut Informatyki Uniwersytetu Warszawskiego; http://www.mimuw.edu.pl/~kszafran/publikacje/tr226.pdf [dostęp: 9.03.2018].

[8] M. Woliński, *Morfeusz – a Practical Tool for the Morphological Analysis of Polish*, w: *Intelligent Information Processing and Web Mining. Advances in Soft Computing*, red. M.A. Kłopotek i in., Berlin 2006, s. 503–512; http://nlp.ipipan.waw.pl/Bib/woli:06.pdf [dostęp: 9.03.2018].

[9] Por. K. Wolski, *SŁOWNIK JĘZYKA POLSKIEGO przez P. LINDE. Do Redaktora Pamiętnika*, „Pamiętnik Warszawski" 1809, 1, s. 35–83; http://ebuw.uw.edu.pl/publication/100787 [dostęp: 9.03.2018].

słowa pochodne **wciąż drukowane**, nie *a capite*, lecz różniąc się **wersalikami**, znowu każde z objaśnieniem, wyłuszczeniem, tłumaczeniem, i t. d.

Traktowanie jako haseł słownika (ogólniej – wyrażeń hasłowych) napisów wyróżnionych wersalikami jest więc naturalne i powszechne. Jednak, jak zobaczymy, na potrzeby indeksu pojęcie hasła zostało zmienione – wprowadzono istotne rozszerzenia (por. pkt 3.6) i niewielkie ograniczenia (por. pkt 3.10). W tym samym przedruku znajduje się również stwierdzenie (s. 31, wyróżnienia moje):

My z naszej strony zaświadczamy wielką poprawność druku […] W niezmiernej masie tego wszystkiego, co się na jednym takowym arkuszu, ściśle wybitym znajduje, niczego nie brak prócz kropki tu i owdzie opuszczonej […] **W ogólności żałować także należy, że nie było na pogotowiu wersalików znaczonych kropkami lub kreskami, któremi wszystkie początkowe słowa drukowano.**

Wolski komentuje to następująco w przypisie na tej samej stronie (wyróżnienia moje):

Że **–** położone jest dla tego bez kropki, bo jest od peryodu; a **nad większemi literami w drukarni kropek nie było**. […] Z drugiej strony stało się już zadosyć życzeniu recensenta, już nawet teraz i **do wersalików głoski przysposobiono z kreskami, i z kropkami**; daje się to widzieć w Słowniku od połowy artykułu pod literą J. Jest to nowy dowód, jak autor ani kosztów, ani trudów względem dzieła swojego nie oszczędza. Przydać i to należy na usprawiedliwienie autora, że **dotąd nie miały prawie wcale drukarnie w używaniu kropkowanych i kreskowanych wersalików; trzeba było dopiero tworzyć je niejako**.

Marian Ptaszyk[10] przedstawia sprawę następująco (cytując dalej również komentarz Wolskiego):

W połowie 112 arkusza w drugim tomie (prawdopodobnie w marcu 1808 r.) zastosowano po raz pierwszy wersaliki ze znakami diakrytycznymi (s. 889: Ć, Ś, Ź; s. 890: Ń; s. 892: Ż; arkusz 114, s. 904: Ó). Możliwe, że zakupiono je u Breitkopfa. […] Odtąd mimo posiadania kompletu wersalików nieczęsto korzystano z tych ze znakami diakrytycznymi. W żywej paginie nieraz zastępowano je literami bez znaków. Trudno dopatrzeć się w tej praktyce jakiejś zasady. Nie tylko Linde nie był konsekwentny w używaniu wersalików ze znakami diakrytycznymi. W znanej książeczce Onufrego Kopczyńskiego drukowanej

[10] M. Ptaszyk, *Słownik języka polskiego Samuela Bogumiła Lindego*, Toruń 2007, s. 72.

w 1808 r. przez warszawskich pijarów *Poprawa błędów w ustnej i pisanej mowie polskiej* znajdujemy na s. 13 **Zeby**, na s. 21 **Smielsza**, na s. 31 **CZYNNOSC**. Podobnie rzecz się ma w innych drukach pijarskich z pierwszych lat XIX w.

Wszystkie cytowane rozważania dotyczą pierwszego wydania słownika. Mają one jednak zastosowanie również do wydania drugiego, które było podstawą do indeksu *a tergo*. Co więcej, w praktyce problemy nie ograniczają się do diakrytów nad wersalikami, ale dotyczą również ogonków i poprzeczki w literze *Ł*. W konsekwencji sporządzenie listy haseł nie jest zadaniem czysto mechanicznym, bo wymaga interpretacji wieloznacznych napisów.

Zasady redakcyjne indeksu[11] brzmią w punkcie IV:

Pisownia haseł indeksowych zgodna jest ze Słownikiem Lindego. Modernizacje dotyczą tylko dużych i małych liter. Jawne błędy druku w Słowniku Lindego, o których świadczy porządek alfabetu i pisownia cytatów, są w indeksie poprawione.

Pominięte diakryty uznano najwyraźniej za „jawne błędy druku", które nie wymagają komentarza. Nie zawsze jednak właściwa pisownia hasła jest oczywista, niewykluczone są też przypadki, że niezgodność indeksu ze słownikiem jest skutkiem błędu drukarskiego w indeksie lub pomyłki jego autorów. Takim wątpliwym przypadkiem jest np. hasło *spółmódlca* (t. 5, s. 393) zapisane w indeksie jako *spółmodlca*[12].

Warto zwrócić uwagę na cytowany wyżej fragment: „Modernizacje dotyczą tylko dużych i małych liter". Ze względu na występowanie w słowniku nazw własnych (ponad 2 000) jawne ich oznaczenie przez użycie dużej litery jest niewątpliwie bardzo pożyteczne – obejmuje to zarówno wprowadzenie dużej litery do napisów pierwotnie wersalikowych, jak i użycie małej litery w słowach pisanych w słowniku dużą literą. Co do właściwej modernizacji pisowni, to wbrew powyższej deklaracji była ona stosowana, choć niezbyt konsekwentnie – np. *bigoteria* zamiast *bigoterya*.

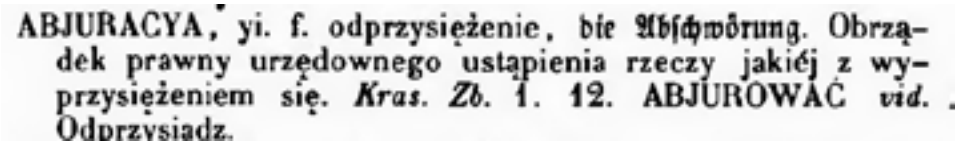
3.2. Hasła wielowyrazowe

W indeksie programowo pomija się „zestawienia" (por. pkt 3.10), ale zostały uwzględnione niektóre wielowyrazowe wyrażenia hasłowe, niekiedy zawierające również znaki interpunkcyjne, np. *da, da, da*. Z techniczne-

[11] R. Grzegorczykowa i in., *op.cit.*, s. 7.
[12] *Ibidem*, s. 14.

go punktu widzenia wielowyrazowe są również te hasła, w których dodatkowe człony są ujęte w nawiasy i mają charakter komentarza, np. *(w) obec*.

3.3. Hasła indywidualne i bloki hasłowe


ABJURACYA, yi. f. odprzysiężenie, die Abſchwörung. Obrządek prawny urzędownego ustąpienia rzeczy jakiéj z wyprzysiężeniem się. *Kras. Zb.* 1. 12. ABJUROWAĆ *vid.* Odprzysiadz.

Rys. 1. Blok hasłowy; w indeksie oba hasła są hasłami podstawowymi

Jak było wspomniane wyżej, „Po pierwiastkowem słowie […] kładą się słowa pochodne, wciąż drukowane, nie a capite". Układ taki, nazywany alfabetyczno-gniazdowym, sprawia problemy terminologiczne. Bilińska opisuje go następująco[13]:

> Hasła w słowniku zostały posortowane *a fronte* w kolejności alfabetycznej, z tym że niektóre zostały uznane za hasła główne (dalej: hasła), a inne, powiązane z głównymi etymologicznie, za hasła niejako podrzędne (dalej: podhasła).

Linde nie uzasadnia nigdzie takiego układu haseł, ale w *Zdaniu sprawy z całego ciągu pracy* cytuje bez komentarza stwierdzenie Wolskiego[14]:

> Z całego układu druku w Słowniku, znać jak starał się ochraniać miejsca; dlatego unikał, ile być mogło, częstych ustępów, *a capite*, i tak daleko ciągnął pasmo słów pochodzących z jednego źrzódła , jak tylko szyk, abecadłowy pozwolił,

tym samym je potwierdzając. Dużą rolę tego czynnika potwierdzają obserwacje Bilińskiej, która pisała[15]:

> Prawdopodobnie też z powodu oszczędności miejsca wiele haseł odsyłaczowych, a więc krótkich, zamieszczono w słowniku nie linia pod linią, a obok siebie […] czy też nawet

[13] J.A. Bilińska, *Analiza i leksykograficzny opis struktury słownika Lindego na potrzeby digitalizacji*, niepublikowana praca doktorska, 2013, s. 74, Wydział Neofilologii Uniwersytetu Warszawskiego; https://depotuw.ceon.pl/handle/item/349 [dostęp: 24.05.2018].

[14] K. Wolski, *op.cit.*, w przedruku s. 32.

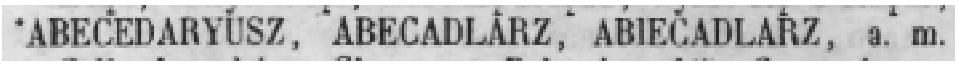
[15] J.A. Bilińska, *op.cit.*, s. 76.

w tym samym bloku, co kolejne hasło […] Zdarzył się też co najmniej jeden homonim zapisany w poprzednim artykule hasłowym […],

ilustrując to m.in. hasłami CZEDŁ, CZEGLANY (t. 1, s. 360) i CZERŃ (s. 364).

Ponieważ terminy *hasło* i *podhasło* mają w indeksie *a tergo* istotnie inne znaczenie, wolimy ich nie używać. W razie potrzeby będziemy mówić o bloku hasłowym, który składa się z *haseł indywidualnych* – a dokładniej z indywidualnych artykułów hasłowych. W indeksie *a tergo* wszystkie hasła indywidualne są traktowane równorzędnie.

3.4. Hasła podstawowe i hasła poboczne



Rys. 2. ABECEDARYUSZ i ABECADLARZ – hasła podstawowe, ABIECADLARZ – hasło poboczne (podhasło)

Hasłami podstawowymi i *hasłami pobocznymi* nazywam jednostki określane w indeksie *a tergo* jako hasła i podhasła. W tym punkcie *hasło* rozumiemy wąsko i technicznie jako wyraz lub wyrażenie hasłowe. Typowy artykuł hasłowy rozpoczyna się kilkoma hasłami i ich kolejność została uznana przez autorów indeksu *a tergo* za tak ważną, że stała się podstawą podziału na *hasła* i *podhasła* indeksowe. Uporządkowane *a tergo* są tylko hasła, a podhasła są składane mniejszym stopniem pisma bezpośrednio pod odpowiednim hasłem.

Merytorycznie podział ten jest uzasadniony tym, że hasła podane w drugiej kolejności to – przynajmniej w zasadzie – warianty fonetyczne lub pisowniowe, np. *cekauz* (hasło) i *cejkhauz*, *cejkauz*, *cajghaus*, *cejghauz*, *cegauz*, *ceghauz*, *cekhauz*, *czekhauz*, *czekauz* (podhasła), *księga* (hasło) i *xięga* (podhasło).

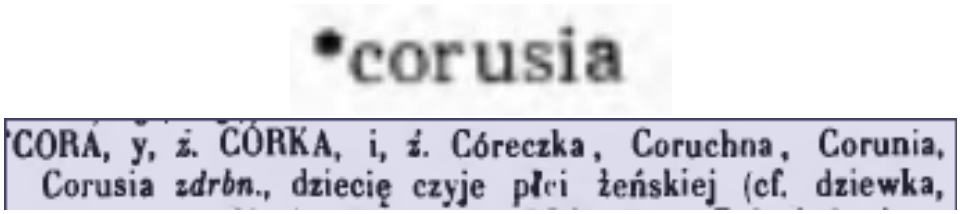
Reguły rozróżniania haseł i podhaseł zostały omówione w punkcie II zasad redakcyjnych indeksu[16], przytaczane tam przykłady są jednak – co łatwo sprawdzić – często niewłaściwe. Na przykład na s. 5 czytamy: „Jako podhasła zostały potraktowane: […] formy typu *spalszczać* w stosunku do *spolszczać* […]”, jednak haseł *spalszczać* i *spolszczać* nie ma ani w słowniku, ani w indeksie.

[16] R. Grzegorczykowa i in., *op.cit.*, s. 4–6.

3.5. Homonimy

Występujące w słowniku numery homonimów przeważnie zostały zachowane, np. *rola 1,2*. Jeśli podhasło odnosi się tylko do jednego z homonimów, jest to zaznaczone jawnie, np. *ad 1 zamięszka* lub *leda ad 2*.

3.6. Hasła wewnętrzne jawne i niejawne



Rys. 3. Przykład jawnego hasła wewnętrznego (hasło indeksowe z gwiazdką)

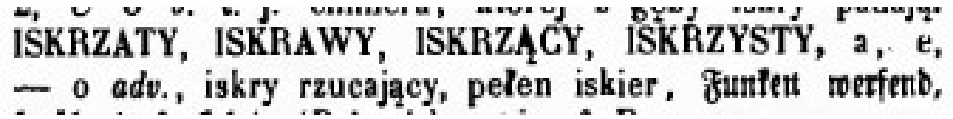
Hasłami wewnętrznymi nazywam hasła oznaczone w indeksie gwiazdką (hasła z gwiazdką występują również w słowniku, ale ma ona tam zupełnie inne znaczenie). Hasła te nie są w słowniku wyróżnione typograficznie.

Zasady redakcyjne indeksu[17] brzmią w punkcie V:

> Hasła opatrzone w indeksie gwiazdką oznaczają wyrazy występujące u Lindego nie w haśle, ale wewnątrz artykułu hasłowego. W ten sposób zostały wydobyte ze Słownika formy niedokonane i częstotliwe czasowników (np. **porębować* pod *porąbić*) oraz przysłówki. […] W wypadku, gdy forma z gwiazdką występuje u Lindego pod hasłem nieoczekiwanym (np. *łacwiusieńko* pod *łacniuchny*, *iskrząco* pod *iskrzaty*) lub takim hasłem, które fonetycznie bardzo się różni od wyrazu poszukiwanego, pod hasłem z gwiazdką umieszczony jest odsyłacz do hasła, w którego artykule forma ta się znajduje.

Podane przykłady wymagają komentarza. Ani w indeksie, ani w słowniku nie ma hasła *łacwiusieńko* (jest to chyba błąd drukarski indeksu), w słowniku nad literą *c* jest krótka pozioma kreska, więc występujące w indeksie na s. 277 odczytanie *łaćwiusieńko* należy uznać za prawidłowe. Tak czy inaczej tego typu hasła nazywamy *hasłami wewnętrznymi jawnymi*. Inna jest sytuacja w przypadku słowa *iskrząco*, które w słowniku w ogóle nie występuje. Zostało ono utworzone przez autorów indeksu na podstawie skróconego i nie do końca jednoznacznego zapisu (patrz rys. 4). Takie hasła nazywamy *hasłami wewnętrznymi niejawnymi*.

[17] *Ibidem*, s. 7.


ISKRZATY, ISKRAWY, ISKRZĄCY, ISKRZYSTY, a, e,
— o *adv.*, iskry rzucający, pełen iskier, Funken werfend,

Rys. 4. Skrótowa reprezentacja przysłówków – hasła wewnętrzne niejawne

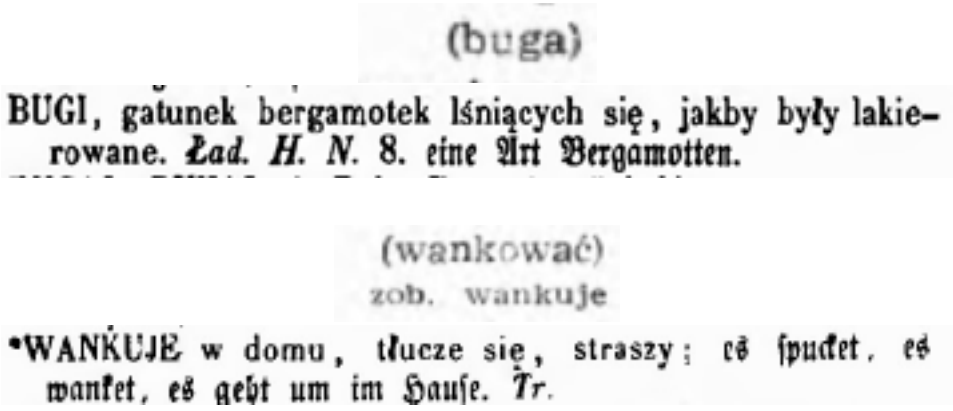
Wspomniany w cytacie odsyłacz ma formę dodatkowego wiersza *zob. iskrzaty*. W indeksie zdarzają sie również pomyłki, kiedy gwiazdka jest – jak się wydaje – mechanicznie przeniesiona ze słownika, np. *wyszpacać*.

3.7. Hasła odtworzone

W punkcie III zasad redakcyjnych (s. 6) czytamy:

Występujące często u Lindego hasła w liczbie mnogiej zachowane są w tej formie (poza wypadkami, kiedy są to niewątpliwe pluralia tantum) wówczas, kiedy rzeczownik w liczbie mnogiej ma inne znaczenie niż w liczbie pojedynczej […] kiedy jest formą równorzędną z liczbą pojedynczą i nie oznacza mnogości […] kiedy wreszcie jest historycznym collectivum: *księża*, *bracia*. W innych wypadkach odtworzona liczba pojedyncza umieszczona jest w nawiasie okrągłym, […] Podobnie w nawiasie rekonstruuje się formę podstawową (M. 1. poj. r. m.) przymiotnika zacytowanego u Lindego tylko w rodzaju żeńskim, nijakim lub liczbie mnogiej, np. *miodonośne* zmieniamy na *miodonośny* chyba że przymiotnik używany jest tylko w rodz. ż.

Przykłady ilustrujące te zasady nie są niestety właściwe, gdyż przytaczanych haseł nie ma w słowniku lub indeksie.


(buga)

BUGI, gatunek bergamotek lśniących się, jakby były lakierowane. *Lad. H. N.* 8. eine Art Bergamotten.

(wankować)

zob. wankuje

•WANKUJE w domu, tłucze się, straszy; es ſpuckt, es wanket, es geht um im Hauſe. *Tr.*

Rys. 5. Hasła odtworzone

3.8. Hasła uzupełniające

W punkcie II zasad redakcyjnych czytamy m.in. (s. 6):

> Hasła potraktowane przez Lindego jako hasła odesłane stanowią w indeksie podhasła, hasła zaś, do których się odsyła, są hasłami głównymi. Zasady tej przestrzega się i wtedy, kiedy Linde odsyła do jakiejś formy podstawowej, której jednak omyłkowo później we właściwym miejscu nie podaje, np. *przeczera zob. przechéra*; mimo że u Lindego forma *przechéra* nie występuje jako hasło główne, w indeksie jest ona uznana za hasło zasadnicze.

W rzeczywistości cytowany odsyłacz (t. 4, s. 512) ma postać: PRZECZERA. *Dudz*. 54, ob. Przechera, ale występuje również odsyłacz (t. 4, s. 506): PRZECHYRA, PRZECHYRNY, *ob*. Przechéra. Hasła *przechéra* rzeczywiście w słowniku nie ma, ale dodanie takiego hasła do indeksu jest w pełni uzasadnione. Hasła takie proponuję nazywać *uzupełniającymi*.

3.9. Hasła pomocnicze

Hasłami pomocniczymi nazywamy hasła mające charakter odsyłacza lub komentarza. W pkt 3.6 podany jest przykład odsyłacza *zob. iskrzaty*, w pkt 3.5 przykład odsyłaczy-komentarzy *ad 1 zamięszka* i *leda ad 2*. Hasła pomocnicze powinny występować w indeksie również jako hasła główne, co pozwala na dodatkową kontrolę poprawności wyników rozpoznawania znaków. Nietypowy charakter ma hasło *(pod-zierać)* stanowiące komentarz do poprzedniego hasła i objaśniające jego wymowę.

3.10. Hasła nieindeksowane

Ze względu na przeznaczenie indeksu zostały w nim w zasadzie pominięte formy fleksyjne, zaimek *się* przy czasownikach, cząstki morfologiczne i skróty. W punkcie I zasad redakcyjnych czytamy (s. 4), że „pominięto również zestawienia typu: *czarna jagoda*, *biała niedziela*", jednak w słowniku nie ma żadnego z tych zestawień. Są natomiast *babczy czosnek*, *biała głowa*, *jedna jagoda*, *kokowe drzewo*.

4. Hasła w indeksie elektronicznym

Podstawową formą indeksu elektronicznego jest komputerowa baza danych, dlatego niezbędne było ustalenie sposobu jednoznacznej identyfikacji haseł. Przyjęto konwencję identyfikowania hasła przez zestaw następujących liczb:

1. Numer strony w indeksie *a tergo* (od 1 do 392, zapisywany zawsze trzycyfrowo). Hasłom pominiętym w indeksie przypisujemy umownie numer strony 999 – na razie jest tylko jedno takie hasło (*tywon*), ale może z czasem pojawić się ich więcej.

2. Numer łamu (kolumny) w indeksie *a tergo* (od 1 do 3). Hasłom pominiętym w indeksie przypisujemy umownie numer kolumny 0.

3. Numer wiersza w łamie w indeksie *a tergo* (od 1 do 55, zapisywany zawsze dwucyfrowo). Jeśli hasło z powodu jego długości zajmuje dwa wiersze, np. *dziedzicznonajemnie*, jest to numer pierwszego wiersza. Hasłom pominiętym w indeksie przypisujemy umownie numer wiersza równy kolejnemu numerowi dodanego hasła – wspomniany wyżej *tywun* ma więc numer 1.

4. Numer homonimu w sensie indeksu *a tergo* lub 0. Cytowany w pkt 3.5 zapis *rola 1,2* odpowiada dwóm hasłom indeksu, których identyfikatory różnią się właśnie numerem homonimu.

5. Numer wariantu hasła, opisany niżej – obecnie prawie zawsze wartość 0.

6. Numer wersji wariantu hasła, opisany niżej – obecnie prawie zawsze wartość 0.

Indeks elektroniczny nie pomija żadnych haseł z indeksu *a tergo.* Jak zostało to pokazane w pkt 3.1, są to często nie oryginalne zapisy słownikowe, lecz ich odczytania, czasami wątpliwe. Pojęcie wariantu hasła zostało wprowadzone po to, aby umożliwić w przyszłości przechowywanie obu tych informacji.

Przewidujemy następujące przypadki:

1. Hasło słownikowe różni się od (pod)hasła indeksowego szeroko rozumianymi diakrytami – chodzi o pary liter: *ó* i *o*, *e* i *é*, *ś* i *s*, *ę* i *e* itd. Obecnie w indeksie elektronicznym hasło występuje tylko w wersji zgodnej z indeksem *a tergo* (informacja o rozbieżności czasami jest umieszczona w komentarzu), por. np. *chrystobójca* i *kapryśnica*. W przyszłości warto utworzyć dodatkowe warianty haseł co najmniej w przypadkach wątpliwych, np. *spółmodlca* (indeks *a tergo* – wariant 0) i *spółmódlca* (słownik – wariant 1), *jędykuła* (indeks *a tergo* – wariant 0) i *jedykuła* (słownik – wariant 1).

2. W indeksie *a tergo* hasło występuje w pisowni zmodernizowanej. W przyszłości warto utworzyć dodatkowe warianty haseł w pisowni oryginalnej, np. *Apolonia* i *Apollonia* (słownik – wariant 2), *ambasada* i *ambassada* (słownik – wariant 2), *bigoteria* i *bigoterya* (słownik – wariant 2), *solenizantka* i *solennizantka* (słownik – wariant 2).

3. Hasło jest odtworzone w sensie pkt 3.7. W przyszłości warto utworzyć dodatkowe warianty haseł w pisowni oryginalnej, np. dla hasła *pokundź* (indeks *a tergo* – wariant 0) hasło *pokundziowie* (słownik – wariant 3).

4. Hasło w indeksie poprawia ewidentny błąd drukarski, np. *przechadzka* (indeks *a tergo*) i *przechachadzka* (słownik). Na potrzeby automatycznej analizy słownika dobrze jest informację o tym zapisać jawnie – do tego celu rezerwujemy numer wariantu 9.

Wersja wariantu hasła w założeniu ma służyć do odnotowywania poprawek do haseł indeksowych. Obecnie mamy tylko jeden taki przypadek – hasło *skwania* wydaje się albo błędem drukarskim, albo błędnym odczytaniem odpowiedniej fiszki. Jest ono zachowane w indeksie elektronicznym jako wersja 0, ale dodatkowo zostało utworzone hasło poprawione *skwama*, którego identyfikator różni się od identyfikatora hasła błędnego tylko numerem wersji – jest on równy 1.

Traktowanie wieloczłonowych wyrażeń hasłowych w indeksie elektronicznym jest jeszcze sprawą otwartą. Prowizorycznie przyjęto następujące konwencje:

– dla wybranych wyrażeń poszczególne ważniejsze człony hasła otrzymały numer wariantu 4, a numer wersji oznacza w tym przypadku kolejny numer członu w wyrażeniu;

– dla wyrażeń zawierających nawiasy, np. *(w) obec*, utworzono wersje beznawiasowe o numerze 1.

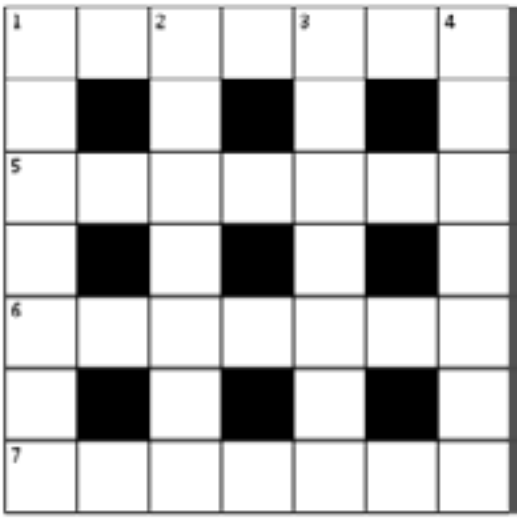
5. Uwagi końcowe

Indeks elektroniczny traktuję jako utwór pochodny w stosunku do indeksu *a tergo* i w związku z tym udostępniam go na identycznej licencji, tzn. *Creative Commons Uznanie autorstwa – Na tych samych warunkach* (por. pkt 2.2). Pliki indeksu razem z dość obszerną dokumentacją są dostępne pod adresem: https://bitbucket.org/jsbien/ilindecsv.

Do wykorzystywania indeksu zgodnie z jego podstawowym przeznaczeniem, to znaczy do przeglądania słownika Lindego, służy program dostępny pod adresem: https://bitbucket.org/mrudolf/djview-poliqarp.

Indeks można jednak wykorzystywać także do różnych innych celów, na przykład dla uatrakcyjnienia zajęć dydaktycznych można przygotowywać krzyżówki za pomocą odpowiedniego programu, np. Qxw. Dzięki udostępnieniu słownika Lindego jako przeszukiwalnego korpusu[18] krzyżówka przedstawiona na rys. 6 może zostać rozwiązana w ciągu minuty – por. rys. 7.

[18] J.S. Bień, *Skanowane teksty jako korpusy*, „Prace Filologiczne" 2012, LXIII, s. 25–36; http://www.ceeol.com/search/article-detail?id=100302 [dostęp: 9.03.2018].



Rys. 6. Krzyżówka stworzona za pomocą programu Qxw na podstawie elektronicznego indeksu. **Poziomo**: 1. miejsce oparkanione; 5. wiara w rzeczy nie godziwe do wierzenia; 6. świadectwa na piśmie; 7. pokrzykanie. **Pionowo**: 1. rów na około czego okopany, lub przez co przebity; 2. liściane ozdoby; 3. biegiem doścignąć; 4. baranek wyrobiony z wosku święconego.

Rys. 7. Rozwiązanie krzyżówki

Podziękowanie

Artykuł został pierwotnie przygotowany za pomocą systemu XƎLATEX, a do wymagań Redakcji przystosowała go Joanna Bilińska, która również zasugerowała poprawki stylistyczne.

Posłowie (czerwiec 2017 roku)

Artykuł niniejszy powstał na podstawie referatu wygłoszonego na konferencji *V Glosa do leksykografii*, która odbyła się w dniach 18–19 września 2014 roku w Warszawie (był to referat plenarny inaugurujący konferencję, slajdy są dostępne pod adresem: http://bc.klf.uw.edu.pl/379/). Zgodnie z instrukcją organizatorów w grudniu 2014 roku złożyłem artykuł do druku w „Pracach Filologicznych"; w czerwcu 2017 roku zostałem poinformowany, że artykuł został odrzucony jako niezgodny z profilem tego czasopisma.

Indeks i program do jego obsługi są nadal rozwijane, aktualne wersje można znaleźć pod podanymi wyżej adresami. Są to tzw. repozytoria zawierające w szczególności mniej lub bardziej szczegółowe historie zmian. Tam też można zgłaszać uwagi, błędy i poprawki.

Wspomniana na wstępie dygitalizacja słownika Lindego (z wyszukiwarką) jest obecnie utrzymywana przez Fundację Języka Polskiego pod adresem: https://szukajwslownikach.uw.edu.pl/.

Bibliografia

Bień J.S., *Polskie zasoby językowe w projekcie IMPACT*, 2011; https://www.slideshare.net/jsbien/polskie-zasoby-jzykowe-w-projekcie-impact [dostęp: 24.05.2018].

Bień J.S., *Skanowane teksty jako korpusy*, „Prace Filologiczne" 2012, LXIII, s. 25–36; http://www.ceeol.com/search/article-detail?id=100302 [dostęp: 9.03.2018].

Bilińska J.A., *Analiza i leksykograficzny opis struktury słownika Lindego na potrzeby digitalizacji*, niepublikowana praca doktorska, 2013, Wydział Neofilologii Uniwersytetu Warszawskiego; https://depotuw.ceon.pl/handle/item/349 [dostęp: 24.05.2018].

Grzegorczykowa R. i in., *Indeks a tergo do Słownika języka polskiego S.B. Lindego*, red. W. Doroszewski, 1965; http://ebuw.uw.edu.pl/publication/339849 [dostęp: 9.03.2018].

Horbaczewski R., *Nagłówki przepisów muszą być różne*, „Rzeczpospolita" 2006; http://archiwum.rp.pl/artykul/607574-Naglowki-przepisow-musza-byc-rozne.html [dostęp: 9.03.2018].

Olejniczak T., *Obsługa formatu PDF/A na potrzeby dygitalizacji tekstów*, niepublikowana praca magisterska, 2011, Wydział Matematyki i Informatyki Uniwersytetu Warszawskiego; https://bitbucket.org/jsbien/pdfautils-fork/downloads/mgr_to236111.pdf [dostęp: 24.05.2018].

Ptaszyk M., *Słownik języka polskiego Samuela Bogumiła Lindego*, Toruń 2007.

Szafran K., *Analizator morfologiczny SAM-95: opis użytkowy*. TR 96-05 (226), Warszawa 1996, Instytut Informatyki Uniwersytetu Warszawskiego; http://www.mimuw.edu.pl/~kszafran/publikacje/tr226.pdf [dostęp: 9.03.2018].

Woliński M., *Morfeusz – a Practical Tool for the Morphological Analysis of Polish*, w: *Intelligent Information Processing and Web Mining. Advances in Soft Computing*, red. M.A. Kłopotek i in., Berlin 2006, s. 503–512; http://nlp.ipipan.waw.pl/Bib/woli:06.pdf [dostęp: 9.03.2018].

Wolski K., *SŁOWNIK JĘZYKA POLSKIEGO przez P. LINDE. Do Redaktora Pamiętnika*, „Pamiętnik Warszawski" 1809, 1, s. 35–83; http://ebuw.uw.edu.pl/publication/100787 [dostęp: 9.03.2018].

An electronic index to Linde's dictionary

SUMMARY

The primary purpose of the index is to facilitate browsing the digitized version of Linde's dictionary. It is based on the reverse index published in 1965, which also has been digitized. Both works are available on the principles of the CC-BY license. The paper discusses the various kind of dictionary and index entries and their representation in the electronic version.

Key words: Samuel Bogumił Linde, dictionary, index, digitization, lexicography.

O Autorze

Janusz S. Bień - profesor zwyczajny w Katedrze Lingwistyki Formalnej Uniwersytetu Warszawskiego, informatyk i lingwista (z wykształcenia matematyk); aktualne zainteresowania to dygitalizacja dawnych tekstów polskich, w tym słownika Lindego i traktatu Parkosza, a także historia pisowni polskiej. Kierował m.in. projektem „Narzędzia dygitalizacji tekstów na potrzeby badań filologicznych” i brał udział w europejskim projekcie „IMPACT - IMProving ACcess to Text”. Wcześniej zajmował się m.in. automatyczną analizą składniową języka polskiego i formalnym aparatem pojęciowym morfologii polskiej.
E-mail: jsbien@mimuw.edu.pl