

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ  
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ  
им. М. Т. КАЛАШНИКОВА

# **Информационные технологии и письменное наследие**

El'Manuscript-2012

Материалы IV международной научной конференции  
Петрозаводск, 3–8 сентября 2012 года

Петрозаводск, Ижевск  
2012

УДК 004.9  
ББК 81.11+81.2-0  
И741

Изданы при поддержке гранта РФФИ (проект № 12-06-06061-г),  
гранта РГНФ (проект № 12-04-14154-г) и в рамках реализации комплекса  
мероприятий Программы стратегического развития ПетрГУ на 2012-2016 г.

Ответственные редакторы:  
В. А. Баранов, д-р филол. наук, проф.  
А. Г. Варфоломеев, канд. физ.-мат. наук, доц.

**Информационные технологии и письменное наследие** [Текст] :  
И741 материалы IV междунар. науч. конф. (Петрозаводск, 3–8 сентября  
2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. — Петрозаводск ;  
Ижевск, 2012. — 328 с.

ISBN 978-5-8021-1402-5

Сборник содержит материалы конференции, посвященной современ-  
ным электронным средствам хранения, описания, обработки, исследования  
и публикации памятников письменности и исторических источников.

УДК 004.9  
ББК 81.11+81.2-0

ISBN 978-5-8021-1402-5

© Петрозаводский государственный  
университет, 2012  
© Ижевский государственный технический  
университет им. М. Т. Калашникова, 2012

**СОСТАВЛЕНИЕ ПЕРЕЧНЯ СОКРАЩЕННЫХ НАЗВАНИЙ ЯЗЫКОВ В  
РАМКАХ ПРОЕКТА ДИГИТАЛИЗАЦИИ  
«СЛОВАРЯ ПОЛЬСКОГО ЯЗЫКА» С. Б. ЛИНДЕ**

*Иоанна Билиньска  
Варшавский университет, Варшава, Польша*

The paper is devoted to the digitalisation of Dictionary of Polish by Samuel Bogumił Linde. The author gives a brief description of a new lexicographic browser available at the Warsaw University and analyses the building process and usage of an index of languages' names abbreviations for digitalisation purposes.

Шеститомный «Словарь польского языка» (1807–1814) авторства Самуила Богумила Линде [Linde, 1807] — выдающееся научное достижение своей эпохи, удостоившееся высокой оценки не только в Польше, но и во всем мире. Нынешняя цифровая версия словаря, размещенная, в частности, на Интернет-портале Куявско-Поморской цифровой библиотеки (г.Торунь),<sup>1</sup> входит в число наиболее популярных и часто используемых публикаций. В то же время, в настоящее время она представляет собой практически только фотографии страниц (сканы), что ограничивает возможности для поиска с использованием компьютерных инструментов.

В настоящее время в Варшавском университете продолжается работа по усовершенствованию дигитализированной версии 2-го издания словаря Линде с использованием технологии оптического распознавания текста (OCR), одновременно совмещенная с поисковой системой<sup>2</sup> [Linde WBL]. Новизна этого подхода заключается в отношении к произведению С.Б.Линде именно как к корпусу текстов, а не просто словарю [Bień, 2011]. Корпус, в настоящее время содержащий около 7 миллионов сегментов, был составлен из сканированных файлов, преобразованных в формат DjVu. С целью обработки поисковых запросов была применена программа Poliqarp for DjVu (marasca)<sup>3</sup>. Основой для нее послужила программа Poliqarp, ранее использованная при создании Национального корпуса польского языка<sup>4</sup>.

Poliqarp for DjVu (marasca) имеет целый ряд достоинств. К их числу относится отображение результатов поиска как конкорданций, связанных с файлами и соответствующим образом обозначенных, а также возможность использования регулярных выражений.

В то же время, максимально полное использование возможностей лексикографической поисковой системы и цифровой версии словаря невозможно без подробного описания его структуры, которое отличалось бы от тех, которые были предложены до сих пор. Эта нелегкая задача дополнительно осложняется неоднородностью статей словаря С.Б.Линде, а также тем, что значительная часть его разделов написана с использованием шрифтов, отличных от стандартной латиницы (например, кириллицы или готического шрифта). Автор словаря стремился представить в нем всю лексику польского языка, а также соотнести ее с другими языками, прежде всего всеми славянскими.

Учитывая тот факт, что словарь Линде является, по сути, плодом многолетнего труда одного человека, структура словарных статей в нем не всегда одинакова. В основном она выглядит следующим образом:

- заглавное слово,
- окончания,
- грамматическая информация,
- этимология,
- толкование, дефиниция и синонимы,
- переводы на иностранные языки, начинающиеся с сокращенных названий языков (в случае с текстами на немецком языке, набранными готи-ческим шрифтом, сокращенное название зачастую отсутствует),
- иллюстрации: примеры из литературы со ссылками (иногда также на латыни),
- переносные значения слова и фразеологизмы (также с примерами из литературы со ссылками), начинающиеся с тире и знака параграфа,
- другие однокоренные слова начинаются с нового абзаца (словообразное гнездо).

Одной из особенностей словаря Линде как словаря алфавитно-гнездового является объемность и внушительная длина многих статей. Так, после фразеологизмов автор приводит также производные слова. Для примера приведем статью «*rięć*» (пять), не содержащую описания производных слов.

К сожалению, автор не составил полного перечня использованных условных сокращений (в этом числе языковых), отметив, что последние носят очевидный характер [Linde, 1994: LXIV]. В большинстве примеров это соответствует действительности, однако для автоматических запросов такой перечень необходим. По этой причине нами была составлена сводная таблица сокращений. В качестве отправной точки для проделанной работы был выбран краткий перечень, приведенный Линде в начале словаря. Нами были выписаны данные из части со словами, начинающимися с буквы А и «Правил этимологии» Линде [Linde, 1994: XXIII–LXI]. Далее таблица была дополнена благодаря автоматически полученному списку слов, написанных курсивом и данными, которые случайно встретились в словаре. В результате на сегодняшний день нами составлен перечень, включающий в себя более 200 сокращений для около 80 языков и диалектов. При этом лишь некоторые языки (к примеру, русский) имеют только по одному сокращению. В большинстве случаев каждому языку соответствует не менее двух, а к примеру, в случае с венгерским языком — даже 7 сокращений (Węg., Hung., Hg., Hng., Hungar., Ungar., Ung.).

**PIĘĆ**, iu; *Boh.* pět; *Slov.* pat; *Sorab.* 2. pēsč; *Sorab.* 1. pietž, piacz, piacžo; *Carn.* pět; *Vind.* pet; *Croat., Bosn. et Rag.* pět; *Slav.* pět; *Eccl. et Ross.* пять; *Graec.* πέντε; *Pers.* peng; *Ind.* pantscha; *Germ.* fünf. *Observes Numeralia cardinalia apud veteres generis feminini*; n. p. One pięć głupich panien wzięwszy lampy, nie wzięły z sobą oleju. 1 *Leop. Math.* 25, 3. On co wziął pięć grzywien, robił niemi; tak iż zyskał niemi drugą pięć. 1 *Leop. Math.* 25, 16. — 2. Gadasz ni w pięć, ni w dziewięć. *Zabl. Amf.* 37. bez sensu, bez składu, nie klei się twa mowa; weder gehauen, noch gestochen. Ożenił się ni w pięć, ni w dziewięć, z kobieta, której nikt nie zna. *Teat.* 36. b, 33. *Mon.* 65, 368. *ib.* 73, 278. *Teat.* 6, 46; *Rag.* ni slastan, ni maastan - ladaco; *Rus.* ni k 'selu, ni k 'horodu. *Pim. Kam.* 320. (*Slov.* bere pat za dwanaśt, *deceptor*).  
 Pochodz. *pięćdziesiąt etc., piętnaście etc, piąty, piątek, piątka, popiętny; pięść, piąstka.*

Илл. 1. Пример словарной статьи в словаре Линде

Пользуясь этой таблицей, при помощи одного запроса можно найти все сокращения для данного языка и, таким образом, получить список почти всех появлений слов на этом языке. Для этого надо составить запрос по формуле: X"\" | Y"\" | Z"\".", где X, Y, Z — это сокращения, "\"" означает точку, а символ | — это альтернатива. Так, для венгерского языка это выглядит следующим образом: Węg"\" | Hung"\" | Hg"\" | Hng"\" | Hungar"\" | Ungar"\" | Ung"\". "\*" Ниже приведены результаты данного запроса в форме графических конкорданций.

polszczyzny XVI wieku.  
Wrocław — Warszawa,  
1966-???? (praca w toku)

- M. Samuel Bogumił Linde.  
*Słownik języka polskiego* (2.  
wydanie). Lwów 1854-1861.
- B. Chlebowski, F.  
Sulimierski, W. Walewski  
(red.), *Słownik Geograficzny*  
Królestwa Polskiego i innych  
krajów słowiańskich,  
Warszawa 1880-1902

Ustawienia  
Podstawowa pomoc

1.	bol. bob, Oberdeutsch bobn, Hung. bab, Lith. pupa. —	Zakładka
2.	Croat. medved, Hung. medve, Etymon. miód, Ross. u	Zakładka
3.	raax, Hung. ros, Vallis. rhyg, Lat. med. re	Zakładka
4.	epa, Graec. épa, Lat. rapa, Hung. repa, Ital. rava, rava, Gall. ra	Zakładka
5.	радъ, Vind. red, Croat. red, Hung. read, Scec. rad, Lith. redas,	Zakładka
6.	redkou, redkva, Croat. rotkva, Hung. retck, Germ. Rettich, Graecolat.	Zakładka
7.	ъ, Vind. shnura, Croat. snora, Hung. sinor, Germ. Schnur, Kieberjäh	Zakładka
8.	Sorab. eyda, Lapp. atahie, atye, Hung. atya, Turc. ata, Roman. atta,	Zakładka
9.	krzjž, Vind. krishi, Croat. kris, Hung. kereszt, Ross. speerz,	Zakładka
10.	Gr. sakkos, Lat. saccus, Hung. zsák, Lapp. tsakes, Hsp. Gr.	Zakładka
11.	Hsp. Gr. sakkos, Lat. saccus, Hung. zsák, Lapp. tsakes, Hsp.	Zakładka
12.	, Scec. aeria, obsol. Gall. errer, Hung. Statute vorfommt, Ross. pachat	Zakładka
13.	, Scec. aeria, obsol. Gall. errer. Hung. aratni, Lat.	Zakładka
14.	um, Lat. Ital. Germ. Dalmaticae et Hung. Venet. 1595. 4. • Etymo	Zakładka
15.	L, Dalmat., Boh., Pol., Ger. et Hung. — studio • Grammatica S	Zakładka
16.	elo Bellosztenécza, nad to Weg., Hung., Hg., Węgierskie słowa	Zakładka
17.	lloszténécza, nad to Weg., Hung., Hg., Węgierskie słowa te ty	Zakładka
18.	tenécza, nad to Weg., Hung., Hg., Węgierskie słowa te tylko u	Zakładka
19.	na Teofila w Wiln. 1594. 4. Weg. Węgierskiego organy. 4.	Zakładka
20.	b ledwo Lat. at, atque, et Hung. és, is, Ross. да; cf. Polon. d:	Zakładka
21.	Sorab. inf. - l - Dalm. alli; Hung. avvagy; Bozn. il, illi, alli; Ra:	Zakładka
22.	se. Boh. dumassek; Croat et Hung. damask; Vind. tomosbk;	Zakładka
23.	ADWENT, u. m. Boh. Slovae. Ungur. idem. Ross. рождество- Afel	Zakładka

## Илл. 2. Результат поиска всех сокращений для венгерского языка

Кроме таких появлений слов различных языков, они также встречались записанными с информацией типа «у россиян», «у чехов», «по-польски», «в венгерском», «с греческого». На следующих иллюстрациях можно увидеть результаты поисков:

- u .\*n — для фраз типа «у россиян»,
- (po|Po)(.\*sku|. \*ску)) — для фраз типа «по-польски»,
- z .\*iego — для фраз типа «с греческого».

Zapytanie: u.\*n

Szukaj

## Wyniki

Znaleziono 180 wyników

Wyświetlanie wyników 1-25

1.	żny), u Rosssyan рѣзкіи, рѣзокъ znaczy rażąc	Zakładka
2.	wyberam, will, wáble; u Rosssyan woliu znaczy chce, ży- bec	Zakładka
3.	veda chłodnik, kuczka letnia; u Rosssyan besieda Mundarten richten. Be	Zakładka
4.	Chytry u Rosssyan i Czechów toż samo co u	Zakładka
5.	Pan; u Rosssyan gospod (cf. gospodarz); u Sor	Zakładka
6.	y u Czechów to co i u nas; u Rosssyan zaś upria- dieß lehtë bedeute	Zakładka
7.	Narzędzie u nas instrument; u Rosssyan Na- den Polen serce; бeрм	Zakładka
8.	ment lub naczynie nadoba; u Rosssyan nadobe Benden kniez (der Ji	Zakładka
9.	m przywarę, i t. d. Gramota u Rosssyan dy- sterfrau, die Ronne,	Zakładka
10.	bey den Croaten zählen, rechnen, und broi die Rabi, die	Zakładka
11.	tyków szczególnie imieniny, u Rosssyan god Ewigfeit, погода гунст	Zakładka
12.	o się, nie tylko u nas, lecz i u Russen god Jahr, godina Zeit, Год	Zakładka
13.	u Rosssyan, Krainczyków Windów, i t. c	Zakładka
14.	u Polaków służy szczekać, u Rosssyan, łajal, Lat. na. — Von dem	Zakładka
15.	gędziec, Boh. hudec, muzyk, u Win- den godeti, die Violin spielen,	Zakładka
16.	te słowa trwają; u Rosssyan zmieniło się tylko hudił,	Zakładka
17.	Abwaschung. Umywanie ciała u pogan obrządek duchowny,	Zakładka
18.	do konfederacyi. 2., przy- u Rzymian na cześć bogów albo wielki	Zakładka

Илл. 3. Результаты поиска для фраз типа «у россиян»

Zapytanie: (po|Po)(.\*sku|.cku)

Szukaj

## Wyniki

Znaleziono 554 wyniki

Wyświetlanie wyników 176-200

Poprzednich 25

176.	v prawie też sa- wie się po Niemiecku Stußofen, co się znaczy pi	Zakładka
177.	albo <i>intestinum coecum</i> , po Polsku kątlica. <i>Urzed.</i> 30. — 98i	Zakładka
178.	<i>Post. W.</i> 2, 77. 'Katolicko, po katolicku <i>adv.</i> , TOWCZYK, a. m.,	Zakładka
179.	n. p. Po katolicku mówiac. <i>Torz.</i> 276. Niekato	Zakładka
180.	z. f. (t. j. pozabijał ich). Po katowsku, *katowsko	Zakładka
181.	zaczął zwadę, Ale po kawalersku, ale nie przez zdradę. <i>KA</i>	Zakładka
182.	<i>rdz. Trag.</i> 197. (meźnie). Po kawalersku wytne się Kawecan dl	Zakładka
183.	ją letko po kawalersku. <i>Zabł. Roz.</i> 119. (z pańska)	Zakładka
184.	władzą królewską zagajony, co po Polsku snadniej	Zakładka
185.	mi widzę zaczynasz prawie po kaznodziejsku. <i>Teat.</i> 52. d, KCIEĆ, kei	Zakładka
186.	ozwo- w człowieku <i>itium</i> , po Polsku kielbaśnica. <i>Urzed.</i> 30. ber	Zakładka
187.	mek i miasto w Kurlandyi, po Polsku Kiesią nazywają.	Zakładka
188.	po rusku kwacz. <i>Włod., Dudz.</i> 41. ber	Zakładka
189.	żnaby niemal powiedzieć, że po Polsku jest to	Zakładka
190.	отворогъ. Suknia zrobiona po Angielsku,	Zakładka

Илл. 4. Результаты поиска для фраз типа «по-польски»

Zapytanie: z.\*iego

Szukaj

## Wyniki

Znaleziono 585 wyników

Wyświetlanie wyników 26-50

Poprzednich 25

26.	sło nie wszczyna się z drugiego ramię, i nie są w obudwóch	Zakładka
27.	ten kształt i u nas <i>śpital</i> , z Łacińskiego <i>Hospit-</i> rиб Pol. grzyb; no	Zakładka
28.	dokładniej wyprowadzać, jak z Rossyjskiego <i>Korowod</i> , choro- Kóunte m	Zakładka
29.	kiego <i>Śchlacht</i> bitwa; lecz z Niemieckiego <i>Geschlacht</i> , wy- (man v	Zakładka
30.	; że zaś tkwi w pochodzących z niego, <i>Jahr verlieben. Erst wenn ma</i>	Zakładka
31.	wyraźnie, istotne głoski źródłosłowa swego zachow zaśm wuwód ich z niego choć mniei oczwivistv na wi	Zakładka
32.	pochodzących. Tak z Greckiego, <i>stan</i> , <i>stenai</i> (Lat. <i>stare</i> ,	Zakładka
33.	mających się tegoż, wyszły z Greckiego języka n. p. mo- nał, amb	Zakładka

Илл. 5. Результаты поиска для фраз типа «с греческого»

Поисковая система с использованием регулярных выражений делает возможным поиск в словаре значительного количества информации. В то же время, однако, можно предположить, что неспециалист, не имеющий навыков работы с дополнительными перечнями и инструкциями, может по-прежнему испытывать определенные трудности в работе с системой. В связи с этим, положительное значение имело бы включение непосредственно в корпус таких сведений, как развитие сокращений (например, в качестве новых тэгов, считываемых программой Poliqarp).

С точки зрения удобства пользователя, интерес представляла бы также разработка возможности появления в тексте словаря развития аббревиатур при подведении курсора. В качестве примера приведем фрагмент «Словаря польского языка XVI века» [SJP XVI] описанный с применением программы студента Норберта Потоцкого<sup>1</sup>:

I. W funkcji przestrzennej (1224):  
 A. *Bardzo blisko, w bezpośrednim sąsiedztwie, obok czegoś; w pobliżu, w niedużej odległości [obiekty stykają się ze sobą lub jest między nimi niewielki odstęp]; ad Miech, Vulg, PolAnt, Mącz, Modrz, JanStat, Cn; iuxta Murm, Miech, Vulg, PolAnt, JanStat; secus Murm, Miech, PolAnt, Vulg, Mącz; circa Miech, Vulg, Mącz, JanStat; secundum Miech, Cn; in, super Vulg, PolAnt; prope Vulg, JanPrzyw; apud Mącz, PolAnt; sub Mącz, Cn; contra, supra, infra, An; praeter, propter Mącz (1284):*  
 AA. *O obiektach nienaruszających się (lub od których jest nieis-*

zad do białochwałitwa vdañ [...]. a przy wyłokim drzewie/ które drzewo łowilżá wowáli. *SkarŻyw* 517 (9); *MWilKHist* A widziano ná Niebie przy okregu Mieñi łpolnie fię bijących *StryjKron* 532 (2); *Cz* iąc iáko przy mocnym murze/ przy piñi kom náłzym/ zářáwáwřzy fię/ co newni M. Miechowita, *Descriptio Sarmatianum Asia-nae et Europianae*, Kraków 1521. 37. patrz góy wieniec wjga ze polpolić *KochFr* 131 (4); *Chćielifmy* twe geñli [J náłterłkiemu Bogu Iti. *Anollinowil* Ná

## Илл. 6. Вид развития для аббревиатуры

Улучшенная таким образом версия корпуса текстов (которой в настоящее время является словарь Линде), будучи объединена со специализированной поисковой системой, может использоваться исследователями, обрабатывающими исторические словари с целью сверки сносок к словарю С.Б.Линде. Она, несомненно, также заинтересует специалистов-филологов, лиц, изучающих лексику, историю и культуру этого периода, а также лексикографов. В то же время, многоязычный словарь С.Б.Линде может использоваться не только филологами-полонистами, но и другими славистами. Неслучайно в прошлом словарь оказал влияние на лексикографию других народов, например словарь чешского языка Й.Юнгмана.

Качественная дигитализированная версия словаря с правильно распознанным или профессионально описанным текстом с применением тэгов является намного более совершенным исследовательским инструментом, нежели его бумажная (или сканированная) версия, предоставляя возможности поиска и копирования соответствующих фрагментов. Кроме того, подготовка соответствующим образом описанных частных элементов словаря дала бы возможность разработки различных графических версий словаря, а также возможности скрытия или предоставления некоторых типов информации, в зависимости от запросов пользователя.

## Список литературы

Bieñ, 2011 — Janusz S. Bieñ, Efficient search in hidden text of large DjVu documents. In: *Advanced Language Technologies for Digital Libraries. Lecture Notes in Computer Science*. Vol. 6699. Springer, Pp. 1–14, <http://bc.klf.uw.edu.pl/177/>.

Linde, 1807 — Samuel B. Linde, *Słownik języka polskiego*, изд. I, Warszawa, 1807–1814.

Linde, 1854 — Samuel B. Linde, *Słownik języka polskiego*, изд. II, Zakład Narodowy im. Ossolińskich, Lwów, 1854–1860, <http://kpbc.umk.pl/publication/8173>.

Linde, 1994 — Samuel B. Linde, *Słownik języka polskiego*, изд. II, репринт, Wydawnictwo Gutenberg-Print, Warszawa, 1994, <http://kpbc.umk.pl/publication/8173>.

Linde WBL — Samuel Bogumił Linde, *Słownik języka polskiego*, оцифрованная версия с поисковой системой, <http://poliqarp.wbl.klf.uw.edu.pl/extra/linde/index.djvu>.

Matuszczyk, 2006 — Bożena Matuszczyk, Słownik języka polskiego S. B. Lindego. Warsztat leksykografa, Wydawnictwo KUL, Lublin, 2006.  
SJP XVI — Słownik polszczyzny XVI wieku, Polska Akademia Nauk. Instytut Badań Literackich, Zakład Narodowy im. Ossolińskich, Wrocław, 1966, <http://www.spxvi.edu.pl/spxvi/slownik.php?s=2>.