Describing Linde's Dictionary of Polish for Digitalisation Purposes

Joanna Bilińska

Formal Linguistics Department, University of Warsaw, Browarna 8/10, 00-927 Warsaw, Poland E-mail: j.bilinska@uw.edu.pl

Abstract

The present paper describes the attempts at digitalising the so called Linde's dictionary of Polish published in 6 volumes between 1807 and 1814 by Samuel Bogumił Linde. We are working on a formal description of the dictionary's structure, whose purpose will be to allow programmers to design a tool for automatic tagging of the text. The dictionary is multilingual, so performing OCR with good quality is a difficult task. The paper also describes the indexes that are going to be added. Compiling an a tergo index and indexes of abbreviations, qualifiers and the names of quotation authors would improve the quality and usefulness of the digitalised version. Our work with the 2nd edition of the dictionary (1854-1861) allows us to test several tools (in different stages of development) that are being developed within the framework of a Polish government grant directed by Janusz S. Bień.

Keywords: digitalisation; old dictionaries; Linde's dictionary of Polish

1. Linde's dictionary

The paper will demonstrate the attempts made to digitalise the so called Linde's dictionary of Polish (Linde, 1807-14) published in 6 volumes between 1807 and 1814 by Samuel Bogumił Linde. It was the first work of such kind for Polish and it met with excellent reception in Poland and abroad. Being a part of Polish cultural heritage, it ought to be represented in digital form to allow more people to get acquainted with it and to enable more advanced usage of it.

The author's intention was to present the Polish language extensively. The dictionary contains as much Polish vocabulary as the author was able to find in available texts. Every word was supplied with all typical pieces of information, such as grammatical properties, definition, quotations from source texts. Moreover, translations are provided into German (in Gothic), Slavic languages, and sometimes also other languages (e.g. Hebrew), as in the author's opinion they were useful to understand older Polish words. The dictionary is both descriptive and normative, because it includes additional information if a word is not used anymore or whether it is more likely to be encountered in poetry or in speech.

MAGIEL, glu, m., et gli, ź., Boh. mandl; Vind. likaunik, likalu, rolovifhe, povalavifhe, povalilu; Ross. KATOKE;
a) narzędzie do gładzenia płótna lub chust pranych, wałkownica, bie Mangel, Manbel, Rolle, bie Błafche su glätten. Jaworowe drzewo do magłów płóciennych bardzo zdatne. Ład. H. N. 54. Magle do chust, warsztat bywa dębowy, wałki dębowe, brzozowe, gdzie zaś płótna bielą, jaworowe. Kluk. Rosl. 2, 161 et 21. Magiel w fabrykach płóciennych składa się z siedmiu ruchomych wałków, pionowo nad sobą ułożonych, w dwóch bełkach osadzonych, nie więcej od siebie odległych, jak tylko, aby płótno między niemi przesunąć się mogło. Przędz. 96, cf. kalendra. — §. b) transl. Daremne były jej bryże, jej maglę. Pot. Syl. 78, t. j. magłowanie, gładzenie twarzy. Pochodz. magłownia, magłować, magłowany, pomagło-

Pochodz. magtownia, magtować, magtowany, pomagtować, wymaglować.

Figure 1: An example entry from Linde's dictionary

Due to its multilingualism, the dictionary's usefulness as a research resources is not only limited to Poland. It could be used for research purposes by any historian, librarian, or lexicographer interested in other (mainly Slavic) languages and cultures. Moreover, it can be interesting for other scholars studying old books, especially dictionaries, and people interested in digitising them.

We are trying to create a formal description of the dictionary's structure, as this could be used by programmers to tag the text automatically in terms of entry names and abbreviations, especially those naming the languages.

2. Digitalisations

Several digital versions of this dictionary exist in digital libraries, for example Google Books and Kujawsko-Pomorska Biblioteka Cyfrowa¹. Both the 1st and the 2nd edition are freely available on the Internet but their quality is not perfect. Generally, they are available in scanned form with OCR that is far from perfect. As such, they are great examples of care of the Polish heritage but at the same time they are useless for research purposes.

However, since the dictionary contains a lot of vocabulary in other languages than Polish, different alphabets and fonts, it is very difficult to perform good OCR. Unfortunately, FineReader does not work sufficiently well with multilingual texts and does not recognise texts that are written in Gothic. While the OCR of Polish parts in the dictionary is good, the parts in other languages are virtually useless.

Furthermore, the book is old and even the 2^{nd} (and last) edition, which was later only reprinted, comes from the 19th century (1854-1861). This results in print errors,

¹http://books.google.pl/books?id=rs0GAAAAQAAJ&printsec= frontcover#v=onepage&q&f=false (Google Books) and http://kpbc.umk.pl/publication/8173

such as variable position of words when they are typed in other alphabets, etc. For example:

boga, von Gott erschaffen. BOGOTÚCZNY, a, e, Eccl. E0готочыт, qui a Deo sunditur, Gr. θεόξοντος, von Gott berfließend. BOGOUBIJCA ob. Bogobójea. **ΒΟΓΟΥΥΠΡ**, a, BOGOWIDZCA, y, m. Eccl. E0говидыць, богозрътель, który boga widzi, θεόπτης, ber Geher Gottes, ber Gott fieht. Bogowidzee Mojżesza usłuchawszy. Smotr. Ap. 1. E0говидънie, вогозрънке, θεοπτία. BOGOWŁADZTWO, a, n teokracya, bogorządztwo, Eccl. богодержавie, Ross.

Figure 2: Variable position of words

Kontekst

, e, Ross. óorocoTBopeHHUH, stworzony od boga, oon ©ott erfóaffen. 'BOGOTOCZNY, a, e, Eccl. K0_ roTOYhtn a Deo funditur, Gr. &eÓQov?og, BOtt ©ott ijerfltefSeuD.' "BOGOURIJCA ob. Bogobójca. BOGOWIDZ, a, BOGOWIDZCA, y, m. Eccl. EoroRHAhi|h, ćoroaptTejb, który boga widzi, &eótirrię, ber ©efyer ©otteg, ber ©ott jietyt. Bogowidzcę Mojżesza usłuchawszy. Smotr. Ap. 1. EoroBHAisHie, Borowie, BOGOWŁADZTWO

Figure 3: OCR quality

This is why this type of digitalisation needs to be done with several specialised tools in addition to a standard OCR program. Some of the tools used will be presented in the paper.

3. Linde's dictionary as a corpus

Searching the hidden text layer in large DjVu files, e.g. dictionaries, is not really efficient as it demands loading the whole file. It was decided that it would be much easier to treat the dictionaries' texts as corpora (Bień, 2011) and to use a specialised search engine for corpora. Therefore, recently a digital version of the 2nd edition of Linde's dictionary was made available at University of Warsaw, with a preliminary OCR (SJPL, 2010). This version of OCR was prepared with FineReader 10 (with 300 dpi resolution) and then converted from PDF to DiVu. It was then converted into a text corpus which consists of ca. million segments now 7 (http://poligarp.wbl.klf.uw.edu.pl/slownik-lindego/).

Interface language:	Query	/: albo
English Change About	Searc	ch
Available texts:		
J. Karłowicz, A. Kryński, W.	Resu	ults
Niedźwiedzki. Dictionary of Polish. Warsaw 1900–1927	Found	423 results so far
 S. Bąk, M. R. Mayenowa, F. Pepłowski (eds.). Dictionary of the 16th century Polish. 	0.000	aying results 1—25
Wrocław — Warszawa, 1966-	1.). Komen- tarz albo wykład na proroctwo Hozeasza,
 ???? (work in progress) M. Samuel Bogumił Linde. 	2.	. Judasz z potomstwem swoim albo kronika ministrów heretyckich przez Przewodowskiego
Dictionary of Polish (2nd	3.	czyli raczej gatunków mowy, albo dyalektów Słowiańskich, niepoślednie miejsce
edition). Lwów 1854-1861.	4.	każdy autor sam siebie, albo też i jeden drugiego objaśnił
 B. Chlebowski, F. Sulimierski, W. Walewski 	5.	się pokazać przed nią? albo : kto ma czoło po
(ed.), The Geographical	6.	się podpisu swego zaprzeć; albo : czoła na to trzeba
Dictionary of the Polish	7.	się pytają. Jabłoń, albo jak Zimorowicz śpiewa: «
Kingdom and other Slavic Countries, Warszawa 1880-	8.	szaty? wnętrzności robaczków, albo skóry zwierząt! Bals.
1902	9.	częstował posły Czeskie dobrze; albo w przysłowie: Ku-
Settings	10.	; ztąd na dobę, albo od doby do do-
 Basic help 	11.	n. p. stnerł albo smart, śmierć. §

Figure 4: Lexicographic browser

The current version of the dictionary can be searched with the Poliqarp for DjVu browser search engine and a concordancer (also called marasca)², which allows users to browse the dictionaries as if they were corpora, returning lists of concordances as search results.

One of the most useful features of the search engine is that the query results can showed as graphical concordances (see figure 5). And as the results are linked with the scans, one can see the original page of the dictionary with the marked result. An example can be seen on figure 6.

Regular expressions can be used for searching, as it is one of the standard features of the Poliqarp concordancer³, which Poliqarp for DjVu is based on. The query syntax is thus the same as in the Poliqarp version used for the National Corpus of Polish, which makes the tool easy to use for people familiar with the earlier version.

Que	ry: Croat	
Sea	arch	
Res	sults	
Four	nd 10000 results	
	laying results 1—25	
Disp	laying results 1-25	
1.		Croat. hlib, ap. Ulphil hlaiban, Anglo
	ыль, долина; Vind. dul, dolina,	
3.		Groat. et Vind. vidjm, vidjti, Lat. via
4	1. Dzień, dnia, Boh. den, dne,	
5.		Groat. denesz, danasz, Lat. hodie), Da
6.	эh. lide Ross. люди, Vind. Iudi,	Croat. lyudi, Germ. Leute, Svec. lyd,
7.	лосъ,	Croat. glasz, Carn. glass, Vind. glass
8.	miasto; ztąd Ross. гражданинъ,	Croat. gradhan . mie- Stadt ; baber Ro.
9.	Stadt; baber Ross. hrazdanin,	Croat. gradhan, ein Burger,
10.	! gnesdu, gnijesdu, gniesdje,	Croat. gnyezdo, Sorab. n'ezdo,
11.	Boh. kwēt, kwjti; Ross. цвътъ,	Croat. czvet, czvit; Vind. zviet, zvie
12.	oudni, pouden, Pol. poludnie,	Croat. Binden in u, j. B. poune, po
13.	sou, Pol. sól; Vind. solitnost,	Croat. poldne, poldan Vind. sol
14.		Croat. medved, Hung. medve, Etymon
15.		Groat. remen, Germ. Riemen, Graec.
		Groat. red, Hung. read, Svec. rad,

Figure 5: Graphical concordances

na u, n. p. poune, pound, pouder, roz potudne, orost. Zanten n u, s. O. poun poldne, poldan. – Vind. Sol, sou, Pol. sol; Vind. solitnost, Croat. poldan, – sounost, Pol. solność, sloność. Sorabowie w Luzacyi na w, Vind. solitnost, sounost, n. p. Zwoto, Pol. złoto; kowo, Pol. koło; swowo, Pol. sło- Die Benben in ber Panij wo (cf. Pol. swoboda, słoboda). U Francuzów odmiana l na kowo, Pol. koło, swowo. u najpospolitsza; z łacińskiego falsus, Gall. faux; felix, boda). Su Straujófijóru if Gall. feu (nieboszczyk); małum, Gall. mal, plur. les maaz. bénig; ans ben Zatinijór Evetfigt), anš małum, nal

§. 25.

§. 20. М. N. Lat. Nicolaus, Ross. Никотай, Pol. Mikolą. Polon. Nicsplik, mesplik, I mesplius, Graec. исопал, Germ. Mispel. — Pol. Nielźwiedz, Boh. nedwed, obs. Boh. Итот. medved, Hung. medve, Elymon. miód, Ross. медь, Boh. med, Lith. meddus, h Weth. Lat. mel, mellis, Graec. ибл. — Polon. niańka cf. mamka.

§. 26.

9. 20. N Dawniejsi wierszopisowie podstawiają często za l, č, Daś N fdyteben bie diten n. p. wzieny, wzieły. – Polon. matżonek, matżonka, matżeń – ny, ftat wzięży. Pol. n stwo, Boh. manżel, manżelka, manżelstwj, Manżonek, mó – Boh. manżel, manżelka, wi Miączyński, niektórzy piszą, dowodząc, iż to słowo "cinige, fagt Miączyński, zdożone jest z Niemieckiego Mann, i z Polskiego żona. "anis ben Deut fd en Wi Budny, tlumacz Biblii, pisze: manżeństwo, manżonkowie. – fammengrięti feu." Der Bi Ktoby chejał wywodzić słowo jelca, (manubrium, rekojeść, stwo, manżonkowie. – 2

Figure 6: Highlighted query result on a DjVu file

For the time being, in the official digital version of Linde's dictionary it is possible to limit the search to individual volumes and the main sections of the dictionary (such as introductory texts, errata, main text, etc.) using the clause "within". Below there is an example of searching for the word "dom" (house) within the 3rd volume.

Interface language:	Que	V: dom within vol3		Query: dom within vol3									
English Change	Search												
About Available texts:	search												
 J. Karłowicz, A. Kryński, W. 	Results												
Niedźwiedzki. Dictionary of Polish. Warsaw 1900–1927	Found 79 results												
 S. Bak, M. R. Mayenowa, F. 		laving results 1—25											
Pepłowski (eds.). Dictionary	Uisp	laying results 1-25											
of the 16th century Polish. Wrocław — Warszawa, 1966-	1.	n n	dom	, rćźni drugie kreskowane,									
???? (work in progress) M. Samuel Bogumił Linde.	2.			zszedł, z Wielopolskich famili									
 M. Samuel Bogumir Linde. Dictionary of Polish (2nd 	3.	, drzewo, materya na											
edition). Lwów 1854-1861.	4.	dół, z drugim w	dom). Zmó- wić									
 B. Chlebowski, F. Sulimierski, W. Walewski 	5.	Ross. sumHKa gliną oblepiony	dom	; Pruski mur; ob									
(ed.), The Geographical	6.	że też meble gubią czasem	dom	cały. Zab. 16									
Dictionary of the Polish	7.	Miasto oppos. stancya,	dom	, n. p .									
Kingdom and other Slavic Countries, Warszawa 1880-	8.	wzgląd na to, żeby	dom	jego nie zginął, i									
1902	9.	. Ty szediszy chłopu w	dom	, wszystko mienisz twoim.									
Settings	10.	nieprzyjmo- wania gości w	dom	przedsięwzięli. Leop. Sap									
Basic help	11.	162. Koń świa-	dom	boju, gdy trąby usłyszy									
	12.	. prebivalische , ztan ,	dom	; Carn. stanuvanje,									
	13.	, stan, kuchia,	dom	; Slav. pribivalac;									

Figure 7: Usage of the "within" clause

4. **Preparing Linde's dictionary for the** users

The dictionary is mainly alphabetical but in the entries one can find also derivates, diminutives, etc. which means that the word order is not strict. Moreover, the Polish characters are ignored which makes the paper dictionary quite difficult to use.

- CALÓWKA, i, 4. lina na cal gruba, ein zolldider Ian. Lina calówka, Gall. alogne, jest długa na 35 sążni, wchodzi całowka, Gali, alogne, jest Gługa na 55 sązni, wchodzi w niej na każde żebro po 72 nitek, waży 100 funtów. Jak. Art. 3, 42. półtora całówka, comblau; ⁵/₃ całówka cinquenelles. ib. 2. Deska całowa, e. jolld. Bret. CAŁOWŁADNOŚĆ, ści, ź. cała władza, pełna moc, naj-wyższa zwierzchność, bie Allgewalt, vóllige Serridajt, Ober-teridati, Cokudatach, pełna moc, naj-wyższa zwierzchność, bie Allgewalt, vóllige Serridajt, Ober-teridati, Cokudatach, pełna moc, naj-wyższa zwierzchność, bie Allgewalt, wollige Serridajt, Ober-teridati, Cokudatach, pełna moc, naj-
- berrichaft. Całowładność narodowa w Polszcze, we trzech stanach zamknieta, na scimie okazuje sie. Skrzet. P. P. 1, 41 et 248. CAŁOWŁADNY, a, e, CAŁOWŁADNIE przystk., nieograniczenie władający, unumfdrauft berrichend.
- CALOWY, a, e, na cal, od cala, 3oll:, einen Bou ober Daumen groß. Ross. дюймовый.
- CAŁOWY, a, e, w cołości czyli całkowitości, na cał, hurtowy, im Ganjen, (en gros). Handel całowy. Gaz. Nar. 1, 115. "CAŁOZUPEŁNY, a, e, z jednej sztuki, burdweg anš einem Etűde. Wędzidła bywaja całozupełne, albo otworzyste. Hipp. 60. CAŁUCHNY, "CZALUCHNY, a, e, CAŁUCZKI, CALUTEŃKI, CAŁUSIEŃKI, CAŁUŚKI, a, ie,

Figure 8: Ordering of lemmas

Considering this, it would be useful to link with it the a tergo index (Grzegorczykowa, 1965), after reverting it to the alphabetical order. It would become a kind of a list of content for the dictionary. The index is already scanned and the OCR was proofread.

http://poligarp.wbl.klf.uw.edu.pl/

³ http://korpus.pl/index.php?lang=en&page=poliqarp

It is important to note that as a result of being multilingual and historic, the dictionary can be valuable especially for Slavonic scholars. For instance, Linde represented the language spoken on the territory of contemporary Slovenia (non-codified at that time) as two separate languages, which he labelled *Crn., Carn. Carniolice* for western dialect and *Vind., Vd.* for eastern dialect. However, the author did not provide a complete list of abbreviations and he used various abbreviations for the same thing (eg. *Boh., Bh.* and *Cz.* for Czech language).

Thus an index of abbreviations is being prepared for the language names. The list could be applied to the OCR version of a dictionary to tag the sections with specific language labels. It could then be used to narrow the search to the specific sections.

At the moment one can simply search the dictionary for the language labels (just typing the abbreviation for the language) and hence get a list of words in this language contained in Linde's dictionary (as the right context in concordances). However, it would be much easier and more comfortable to search within the dictionary parts labelled with language names.

Que	ry: Croat	
Sea	ırch	
Res	sults	
Four	nd 63 results so far	
Disp	laying results 1—25	
	-, , , , , , , , , , , , , , , , , , ,	
1.	, Aliebez, Alebez,	Croat . hiib, ap.
2.	. dul, dolina,	Croat . dól, dolina,
3.	. Biutib, BHŻKY,	Croat . et Vind. vidjm
4.	. den, dne,	Croat . et Yind. dan
5.	, Boh. dnes,	Croat . denesz, danasz,
6.	, Vind. ludi,	Croat . lyudi, Germ.
7.	IO CB,	Croat . glasz, Carn.
8.	ztąd Ross. rpaauaHHHt,	Croat . gradhan = mie-
9.	Ross. /irażdanin,	Croat . gradhan, etn Surger
10.	, gnijesdu, gniesdje,	Croat . gnyezdo, Sorab.
11.	. UB^TŁ,	Croat . czvet, czvit;
12.	, Pol. południe,	Croat . 3Bittben ilt u,
13.	. so/itnost,	Croat . po/dne,
14.	. nedved, medved,	Croat . medved, Hung.
15.	, jermen, jeremen,	Croat . remen, Germ.
16.	, Yind. red,	Croat . red, Hung.

Figure 9: Searching for "Croat" label (Croatian)

It does not seem as if there were any strict rules for the language order within the entries, but we are trying to find regularities in it. With the browser it is possible to search for the language labels and observe whether they appear in the same order all the time. Of course, the same is possible with the paper version of a dictionary, but an electronic version combined with a specialised browser considerably accelerates and facilitates the task.

Preparing a list of other abbreviations is also planned. It could be implemented into the electronic version of the dictionary, enabling the reader to see the meaning of an abbreviation by clicking on it or pointing the mouse on it.

The author did not prepare such a list as he found it obvious and understandable to everybody. Some information on it was published in (Matuszczyk, 2006), but it is not complete and is not aimed to be a list but just a description of the kinds of labels in the dictionary.

Linde's dictionary is well known also because the author included a great number of examples from literature, adding the authors' names as abbreviations. A detailed list of such abbreviations and text titles was provided by the author himself and in the book of Hrabec and Pepłowski (Hrabec & Pepłowski, 1963). It would be very useful if those texts were linked, as there are many researchers working on describing how the texts of the authors they are interested in are exemplified in dictionaries. In the future it would be possible to link also the abbreviations with the digitalised texts in digital libraries.

The abbreviations (language names, quotations authors, qualifiers) are mainly in italics, and once again the prototype of a newer lexicographic browser with font tagging (see page 5) is of a great help, as it allows for searching for a particular font shape.

Furthermore, it would be useful to work out the meaning of punctuation marks which were used by Linde. He used commas, semicolons and ellipsis extensively and generally the usage is clear. However, there are three types of section marks too (two normal and one mirrored), and it would be of much help if we were sure what is the meaning of them. Maybe it is just for the lack of the same fonts but maybe there is a deeper sense.

краска, bte Farbe, bte Couleur. Jakie kwiatki, takie też są barwy. Budn. 4. Ezdr. 9, 17. Radz. ib. Teraz odnawiają lasy swą barwę. Past. F. 8. Straconej nigdy nie odzyska spełna Barwy, gdy farby się napije welna. Zab. 7, 303. Po śmierci Bolesława, wszyscy tak żałośni byli, że nikt w jasnéj barwie nie chodził. Papr. Herb. (wszyscy żałobę nosili). Darwa na twarzy, cera, bue Gefichtšfarbe. Ona twarz 1 barwa zmieniona okazowała wnętrzną boleść serca jego. 1 Leop. 2 Mach. 22, 16. On współ z głosem i barwę stracił, z tej przestrogi. Otw. Ow. 105. Barwa kościelna vid. Cera kościelna. Fig. Znać było po ich twarzach barwę cnoty, gdy zapłonionych uważałem. Mon. 65, 63. (rumieniec, wstyd). -§. Pod barwą, jednej barwy, nie pstry, eutfarbig, niśt bunt. Szata pod barwą, t. j. jednego koloru, jednej maści. Cn. Th. 730. Star. Dw. 29. S. Barwa, liberya, Ross. Impen; Vind. słushaunu oblazhilu, snamski obliek, bte Lipree, (obs. bie Farbe, cf. Spefarbe, Giaatěfarbe).

Figure 10: Two of three types of section marks

5. Tools used for digitising

Several tools were used for the purposes of a new digital version of Linde's dictionary.

For scanning the 2nd edition of Linde's dictionary in 600 dpi resolution we used Scanhelper⁴, a command-line scanning tool prepared by Jakub Wilk. These scans are available at the digital library: http://eprints.wbl.klf.uw.edu.pl/view/creators/Linde=3A Samuel_Bogumi==0142=3A=3A.html.

The preliminary version of the digitalisation which is available for searching with a lexicographic browser (http://poliqarp.wbl.klf.uw.edu.pl/slownik-lindego/)⁵ was prepared in 300 dpi resolution using Kofax VRS software.

The scans were then converted into DjVu which is thought to be the best format for scanned texts, as it allows for instance to store the OCR results in a hidden text layer. Moreover, the documents can be stored as individual pages, which facilitates the usage of the files, especially those containing large texts.

In the framework of the project mentioned earlier, Grzegorz Chimosz prepared a prototype of a DjVu files viewer based on DjView4 which is, at the same time, a Poliqarp client and thus is an alternative for traditional usage of Poliqarp through a web browser.

The results of a query in the form of graphical concordances are shown in the left panel, and in the main panel there is the DjVu file with highlighted results. An example is provided in Figure 11.

When pointing on a word, one can see underneath the hidden text layer (OCR). Moreover, the viewer enables the user to mark the result and copy its URL (Figure 13).

Later, thanks to Tomasz Olejniczak's pdfa2hocr⁶ converter, Linde's dictionary in PDF/A files from FineReader was converted into hOCR files. This was done because Poliqarp for DjVu is now being prepared to enable the user to search for fonts. For this the font-sensitive corpus builder by Marcin Zając (Janusz S. Bień's student) is being used⁷. The program enables converting hOCR files into a corpus for Poliqarp concordancer.

Currently this browser is only used for experiments but it looks very useful for our research purposes. Figures 14, 15, and 16 below show the possible queries and the results.

As you can see on the figures, the OCR program recognizes the little differences between the fonts, although they generally seem the same when looked at by humans and are of no importance within a text. However, with regular expressions one can search for example for all the text typed in a specified font size range (for example all words in font size 12 and more).

Apart from searching for font sizes, users can search for font shape. Figure 16 shows an example of a query for all the words in bold italic.

This feature of the experimental version of the lexicographic browser is of great usefulness for preparing indexes for the dictionary. For example, searching for a text in italic we will get generally language names' abbreviations, qualifiers and quotation authors abbreviations.

⁴ http://jwilk.net/software/scanhelper

⁵ http://poliqarp.wbl.klf.uw.edu.pl/extra/linde/index.djvu

⁶ http://students.mimuw.edu.pl/~to236111/PDFAUtilities/

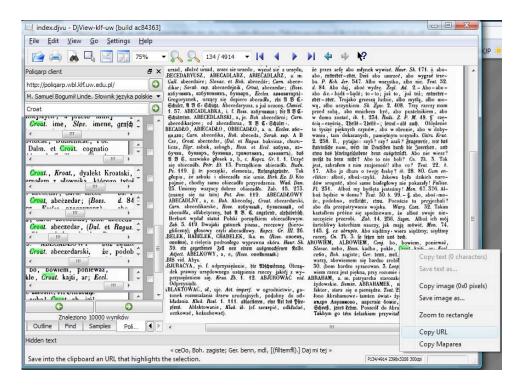
⁷ https://github.com/mzajac/FSCB

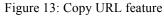
📄 🚔 🚜 🖳 📆 FitWid	dth 🔻 💂	2	82 / 49	14	- 1			M	4	₽	₩?				
	5 × 0	I I słowa sł	Ridre Ridre B. W; (ho cozywiew woje, mół cój postia. Wykanęc, yc. – W kł, powoś do postia. Wykanęc, je postiał. do postiał. na waka postiał. na waka	spidylaža (a f jo jesti b šie pocho sie	i w ktire i w ktire § 16. i w loruj dnych. b i w loruj nie dop zwyczaj, ce, oble roboda, t sko. Ros o się ró rypije. § 17. rzykłady siko, tej ta t renuis ana, Gern d. chych, t renuis anger. –	tive) ja się zs się ja bojow się się so się się so swobodny wobodny s	obą, jaka ad, wojow tedy prz wrad po tedy prz , obyczaj myn. – a becadłe w szyku szy, lepiej dzł, <i>Ial.</i> . Germ. 1 yjne, kor gmiose, z. dyn, – Pel. tać, Rosz skrjaka, 	to w tych aloje, oć svoj, albć svoj, albć vsvoj, albć	. and f. i bellet: . inform . generation of the second . information of the second . I	B-W; olgenben boj (fall), svojna crfinng bo doj (fall), svojna unterfeld balić Oseoc.; nm erf. auferß trijden is, šieo. pol. pige baland tarty, . hidem tarty, . hidem tarty, . hidem tarty,	belder Din (2) bail beili gans of after Simmely, boy a (br: Sim of a Macana after Simmely, boy a (br: Simmely, boy weak	cienki, Bh. ten ill-tendre), Per i, trjti, Boss. 1 Graec. x12liv nja, Germ. Scht skryin, Lett. 2	r e), en in rinanbedisensbedjerden wegtensd (E kelterbetten) (E auch bei persyn auch bei persyn hurcit, obiesid of Obt. en und jolgt ben s. und jolgt ben s. und jolgt ben s. d. – Buchn (Ort gamb), fra in ben 2 key, Base, road of repers, rpy, ar, Let ceck tend, Garant, skryne, Let. L. klob, Alieb	indétheir er- ringera. Strag (elle mad ber n. sun pod- wöj, podböj , odswotny, 6, obied. — ift bir Bigar and in ber ad, puchad, , ef Germ. I. Lat. bibo. Declinationen, mail, Vind. Zalliz. tene, , Unabruck. , Canabruck. med. sere- bez, Alebez,	
Znaleziono 10000 wyników Outline Find Samples Poli	<pre> • • • • • • • • • • • • • • • • • • •</pre>	_		_	_	_	_		_	_	_			_	

Figure 11: Poliqarp client

index.djvu - DjView-klf-uw [build ac84363]	has provide taking in		X
File Edit View Go Settings Help			
100%	• 🕵 💁 134 / 4914 🔹 📢	▲ ▶ ▶ ↓ ↓ ↓ ₩?	
Poligarp client Poligarp client http://poligarp.wbi.klf.uw.edu.pl/ M. Samuel Bogumił Linde. Słownik jezyka polskie Croat Croat Croat prector minicj Croat. Croat. cognatio Croat. Croat. cognatio Croat. Kroat., dyalekt Kroatski, Croat. Croat. cognatio Croat. kaju ar: biological Croat. Acroat. (Bass. d. 84 ± Croat. beczedar; (Bass. d. 84 ± Croat. beczedar; (Cal. et Ragus.) Croat.	 nom einem ausgemachfenen Salbe. ; (Rozz. onofikosuki) iężenie, bie Ubfdmörung. Obrzą- ustapienia rzeczy jakiej z wy- Zb. 1. 12. ABJUROWAC vid. <i>imperf.</i> w ogrodnictwie, ga- urodzajnych, podobny do od- 11. ablaciteren, etne Urt beš Øfro- luk. ub. (ef. szczepić, odkładać, 	wiem rzecz jest piękna, przy rozumie cnota. Rej. Wić. 42. b. ABRAHAM, a. m. patryarcha starozakonny. Wordpam. Imię żydowskie. Demin. ABRAHAMEK, mka. Abrahamek nasz faktor, stara się o pieniądze. Teal. 22. 13. §. fr. Abraham- tono Abrahamowe i tamten świat życie po śmierci, Eecl. wapo Appaswoso, uspcraie Gowie, mećecuoe, %fordpams čdroß, jeneś &tóm. Poszedł do Abrahama, occubuit. Mącz. Takbym go tém żelazkiem przywiał, ażby się z Abraha- III	
	« Slovac. nebo, Bosn. kadbo, pokle,	[Croat.] kajti, ar; Eccl. »	
		P134/4914 2398x3208 300dpi x=1963 y=8	72

Figure 12: Hidden text





Quer	y: albo	
Szul	(aj	
Res	ults	
Foun	d 710 results	
Displ	aying results 1—25	
1.	- riva : abo , albo [:font:straight:bold:TimesNewRoman:8.2pt]	, ali , alić ,
2.	547 . Albo wszystko , albo [:font:straight:bold:TimesNewRoman:8pt]	nic . Teat . 52
3.	Trojako grzeszą ^ Judzie , albo [:font:straight:bold:TimesNewRoman:8pt]	myślą , albo mo -
4.	Judzie , albo myślą , albo [:font:straight:bold:TimesNewRoman:8pt]	mo - wą , albo
5.	albo mo - wą , albo [:font:straight:bold:TimesNewRoman:8pt]	uczynkiem . Sk . Żyw
6.	zabrałem z nim znajomość ! albo [:font:straight:bold:TimesNewRoman:8pt]	co ? Teat . 22
7.	być komu w raju , albo [:font:straight:normal:TimesNewRoman:7.7pt]	w adzie . Smot .
8.	cznym, wpada w afektacya albo [:font:straight:normal:TimesNewRoman:7.7pt]	przesadę . Zab . 4
9.	. Moskwa w swych trebnikach albo [:font:straight:bold:TimesNewRoman:7.8pt]	agiendach . Sak . Pers
10.	m . źle zamiast taiarak albo [:font:straight:bold:TimesNewRoman:7.8pt]	tatarskie ziele . Dodz .
11.	- dno , co procent albo [:font:straight:bold:TimesNewRoman:7.8pt]	prowizya od summy wexlem,
12.	, to jest wino dzikie albo [:font:straight:bold:TimesNewRoman:7.8pt]	leśne . Urs . Gr
13.	Шгфе . Pop wyjmuje hostya albo [:font:straight:bold:TimesNewRoman:7.7pt]	ahnec z proskury , na
14.	tylko stra - cenie roku albo [:font:straight:bold:TimesNewRoman:8pt]	wolności od pozwu , sędzia
15.	^ rocejtemlettung . 0 akcessoryach albo [:font:straight:bold:TimesNewRoman:8pt]	wstępie do spraw » Zachowując
16.	. Akcye poszły w górę albo [:font:straight:bold:TimesNewRoman:8pt]	spadły, tj.

Figure 14: Lexicographic browser with font tagging

Query: [font-size=12.*]

Szukaj

Results

Гания	nd 11 results		
Disp	laying results 1—11		
1.		A, [:font:straight:bold:TrebuchetMS:12pt]	a . A , Głoska
2.	А,	a [:font:straight:bold:TrebuchetMS:12pt]	. A , Głoska najpierwsza
3.	"Cn.Th.	VS [:font:italic:normal:TimesNewRoman:12pt]	. ^ albfeibenjeug . BĘKARTKA
4.	, biednie , ОД ,	Щ [:font:italic:bold:TimesNewRoman:12.497pt]	artnfeelig , elettb
5.	ci - devant) .	B [:font:straight:bold:Tahoma:12.496pt]	z.BZDERE, n
6.	- devant) . B	z [:font:straight:bold:Tahoma:12.496pt]	. BZDERE , n ,
7.	, ша <mark>ф) С</mark> П .	w [:font:straight:normal:Tahoma:12.006pt]	owsach chwaści się śnieć ,
8.	« in feinem \$ erjen	Щ1 [:font:italic:normal:TimesNewRoman:12.001pt]). Serce jej dzwoni
9.	Pot Poez . 147 .	F [:font:straight:bold:TimesNewRoman:12.496pt]	I Ł FRABUGA ob .
10.	Poez . 147 . F	I [:font:straight:bold:TimesNewRoman:12.496pt]	Ł FRABUGA ob . Frambuga
11.	. 147 . F I	Ł [:font:straight:bold:TimesNewRoman:12.496pt]	FRABUGA ob . Frambuga .

Figure 15: Searching for words in font size 12.

 Interface language: 	Query	[font-style=italic & font-weight=bo	old]	
Polish Zmień About	Szuka	i)		
 Available texts: Batch01 Batch01v2 	Resu	lts		
 Batch01v2 Batch01v3 	Found	1000 results		
Batch01v3seg Batch01v4 Batch01_REG_ONLY_final		ying results 1—25		
 Batch01seg_REG_ONLY_final 	1.	. Sk . E . S	Słownik [:font:italic:bold:TimesNewRoman:6.5pt]	Lindego wyd . 2 .
 Wybrane GT_final 	2.	Sk . E . Słownik L	indego [:font:italic:bold:TimesNewRoman:6.5pt]	wyd.2.Tom
 Batch01v4seg 	3.	. E . Słownik Lindego v	wyd [:font:italic:bold:TimesNewRoman:6.5pt]	. 2 . Tom I
 Batch02 	4.	. Słownik Lindego wyd . 2	[:font:italic:bold:TimesNewRoman:6.5pt]	. Tom I . AA
 Batch02v2 Batch02v2seq 	5.	Lindego wyd . 2 . 1	fom [:font:italic:bold:TimesNewRoman:6.5pt]	I . AA - A
 Batch03 	6.	wyd.2.Tom I	[:font:italic:bold:TimesNewRoman:6.5pt]	. AA - A B
 Batch03seg 	7.	, menig * ftenS , m	mm [:font:italic:bold:TimesNewRoman:8.1pt]	аиф nur . Miejcie jeszcze
 Batch04 Batch04seq 	8.	, od stworzenia świata , n	mx [:font:italic:bold:TimesNewRoman:11pt]	2lbam Ijer , wrt (
= Batch05	9.	. 4 , 524 . 5	Stownik [:font:italic:bold:TimesNewRoman:6.2pt]	Lindtgo wyi . t .
 Batch05seg 	10.	4 , 524 . Słownik L	indtgo [:font:italic:bold:TimesNewRoman:6.2pt]	wyi.t.Tom
= Batch06	11.	, 524 . Słownik Lindtgo v	wyi [:font:italic:bold:TimesNewRoman:6.2pt]	.t.Tom I
 Batch06seg Batch07 	12.	. Słownik Lindtgo wyi . t	[:font:italic:bold:TimesNewRoman:6.2pt]	. Tom I . —
 Batch07seg 	13.	. Wyr . SSerglet ^ ii	iber [:font:italic:bold:TimesNewRoman:10.7pt]	ben © eminnfi bee einem
= Batch08	14.	,a.n.i	[:font:italic:bold:TimesNewRoman:7.7pt]	. prawo skarżenia , b
 Batch08seg Batch09 	15.	z Litwy . Nies , i	[:font:italic:bold:TimesNewRoman:10.8pt]	. Kurop . 3,
 Batch09seg 	16.	. Konst . 2 , 1	ii [:font:italic:bold:TimesNewRoman:8pt]	. 2 alić » aż

Figure 16: Searching for words in bold italic

6. Conclusion

As it was mentioned, it is not an easy task to digitise such a dictionary, but we hope that the ongoing formal analysis and lexicographical description of the book will help to better understand the structure of the dictionary and therefore prepare better tools to work with it.

The combination of 1) the indexes to the dictionary that are being prepared with 2) DjVu files of Linde's dictionary and 3) the lexicographic browser with many features valuable for researchers will considerably improve the quality and usefulness of digital dictionary.

While studying the structure of the dictionary (both macro- and microstructure), we are using computer tools already prepared for digitalisation purposes. This both facilitates the analysis, and allows to test the tools and

develop them further. They could be later used for other digitalisation purposes.

In our work we are going to follow the best available examples of the old text digitalisations. We hope that our experience will be of some help for other researchers working on early dictionaries.

7. Acknowledgements

Development of most of the tools described in the present paper is supported by the Polish Ministry of Science and Higher Education's grant no. N N519 384036 Narzędzia dygitalizacji tekstów na potrzeby badań filologicznych directed by Janusz S. Bień. The grant supports also my participation in the conference Electronic lexicography in the 21st century: new applications for new users (eLEX2011).

The paper was proofread by Radosław Moszczyński.

8. References

- Bień, J.S. (2009). Digitalizing dictionaries of Polish. In
 K. Bogacki, J. Cholewa & A. Rozumko (eds.) Methods of Lexical Analysis: Theoretical Assumption and Practical Applications. Wydawnictwo Uniwersytetu w Białymstoku, Białystok, pp. 37-45. Accessed at: http://bc.klf.uw.edu.pl/71/.
- Bień, J.S. (2011). Efficient search in hidden text of large DjVu documents. In *Advanced Language Technologies for Digital Libraries. Lecture Notes in Computer Science (Theoretical Computer Science and General Issues)* (6699). Springer, pp. 1-14. Accessed at: http://bc.klf.uw.edu.pl/177/.
- Grzegorczykowa, R., Kurzowa, Z., Puzynina, J. & Doroszewski, W. (1965). *Indeks a tergo do Słownika języka polskiego S.B. Lindego*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.
- Hrabec, S., Pepłowski, F. (1963). *Wiadomości o autorach i dziełach cytowanych w Słowniku Lindego*. Warszawa : Wiedza Powszechna.
- Linde, S.B. (1807-1814). *Slownik języka polskiego*. 1st ed., Warszawa.
- Linde, S.B. (1854-1860). *Slownik języka polskiego, 2nd ed.* Lwów: Zakład Narodowy im. Ossolińskich. Accessed at: http://kpbc.umk.pl/publication/8173.
- Linde, S.B. (1994). *Słownik języka polskiego*, 2nd ed., *reprint*. Warszawa: Wydawnictwo «Gutenberg-Print». Accessed at: http://kpbc.umk.pl/publication/8173.
- Matuszczyk, B. (2006). *Słownik języka polskiego S.B. Lindego.* Warsztat leksykografa, Lublin: Wydawnictwo KUL.
- Ptaszyk, M. (2007). *Słownik języka polskiego Samuela Bogumiła Lindego*. Szkice bibliologiczne, Toruń: Wydawnictwo Uniwersytetu Mikołaja Kopernika.
- *SJPL* (2010) Linde, S.B., Słownik języka polskiego, digital version under search engine, Accessed at: http://poliqarp.wbl.klf.uw.edu.pl/extra/linde/index.djv u.