Преглед НЦД 6 (2005), 95–102

Tadeusz Piotrowski
(English Department, Opole University)

DIGITIZATION OF POLISH HISTORIC(AL) DICTIONARIES

Abstract: We discus the project of converting Polish dictionaries, both historic and diachronic, into a machine-readable form. It will consist of two stages: one, scanning the dictionaries and storing the images, the other, converting the images to text, by use of OCR programs. The project will be presented on a background of other digitization projects in Poland.
Key words: digitization, dictionaries, Polish lexicography

1. General description. State of the art in Poland.

The paper describes a project which aims at converting the historic and historical dictionaries of Polish into machine-redable form. The project was submitted to the State Committee for Research (Komitet Badań Naukowych, in 2003, it failed then, and was resubmitted in 2004), for funding. The project will cover three multivolume historic dictionaries (one in two editions), from the 19th century, two large historical dictionaries, and a number of one-volume dictionaries. The dictionaries will be described in greater detail in one of the following sections. The books in all have around 55,000 pages.

These dictionaries are historic in the sense that they are monuments to the publishing industry in Poland, to the history of Poland; they were compiled and published at the period of Polish history when Poland was, partitioned, under occupation of three major powers in Europe: Russia, Prussia and Austria. They are also important in that they contain a wealth of linguistic data. One of them was in fact the first diachronic dictionary of Polish, and it contains citiations from publications that no longer exist. These dictionaries, together, still contain the largest amount of data on the Polish lexicon and its development, and their importance has not been superceded by more recent dictionaries. However, these dictionaries cannot be used without difficulty, as they can be found only in research libraries in large cities, even though some of them were re-printed in the 20th c.

Therefore the aim of the project is not only to make images of the dictionaries that would be freely available to the general public, but also to attempt to interpret the images by OCR methods, converting them into text. That is why also the material from the historic dictionaries will be complemented by material from the two major diachronic dictionaries of Polish, Old Polish (Słownik staropolski, 1953–2003; covers the Polish lexicon until the 15th century) and the dictionary of Polish from the 16th century (Słownik polszczyzny XVI wieku 1966–; unfinished), to make available a wide range of data on the Polish lexicon.

For most major European languages there are available digitized historic dictionaries. For English one of the most famous is the monumental Oxford English Dictionary, started in the 19th century, finished in the 20th

(cf. http://dictionary.oed.com/; Weiner 1985), or the Samuel Johnson dictionary from the 18th century (cf. http://www.cblprojects.com/jd/; McDermott 1996, 2003). A very interesting edition, prepared by an amateur without any outside funding, is that of the American Century Dictionary from the 19th c. ("America's greatest dictionary") in DjVu technology (http://www.global-language.com/CENTURY/). There is a number of dictionaries for German, most notably the largest dictionary, started by the Grimm brothers (cf. http://germa83.uni-trier.de/DWB/welcome.htm; cf. Christmann and Schares 2003), or the Middle German dictionaries (http://gaer27.uni-trier.de/MWV-online/MWV-online.html; cf. Burch and Fournier and Gärtner 2002). Most French historic dictionaries are available in a digitized form, and perhaps all editions of Dictionnaire de l'Académie française are available (cf. http://www.chass.utoronto.ca/~wulfric/dico_tactweb/acad.htm). For Slavonic dictionaries, the famous nineteenth century dictionary of Russian by Vladimir Dal' is also available, in several versions and editions (e.g., http://vidahl.agava.ru/, http://www.slova.ru/; interestingly, none has the edition edited by Jan Baudouin de Courtenay, with impolite words), as well as other more modern dictionaries (e.g., http://dic.academic.ru/). For Czech work is under way to digitize the most important dictionaries (Karel Palá, Jana Klimová, personal communication; cf. also Králík and Šmídová 2000).

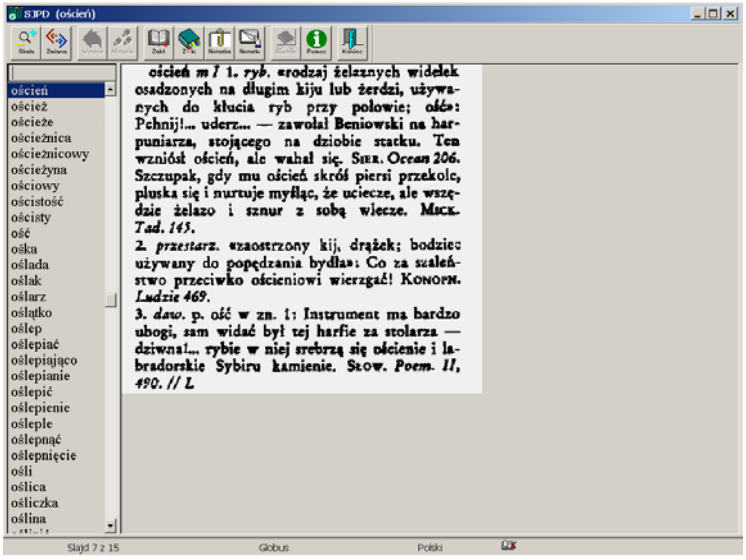


Figure 1 Dictionary by Doroszewski

For Polish unfortunately there are only two large historic dictionaries, one of them encyclopedic, that exist in electronic form. One is Słownik języka polskiego edited by Doroszewski, and published 1958–1969. Most likely this dictionary will not be revised any longer, unfortunately (and there is some controversy over the copyright to it), as it is the largest dictionary with the vocabulary from the 20th c. In the 1990's, on the initiative of Janusz S. Bień, the dictionary was scanned and the TIF images made available on one CD-ROM. The CD includes a search application, which allows the user to go quickly to an individual entry. Unfortunately, the images are rather low in quality (200 dpi), and they are not very clear. There were also attempts to OCR the dictionary, but because of the quality of the scans they failed. The encyclopedic

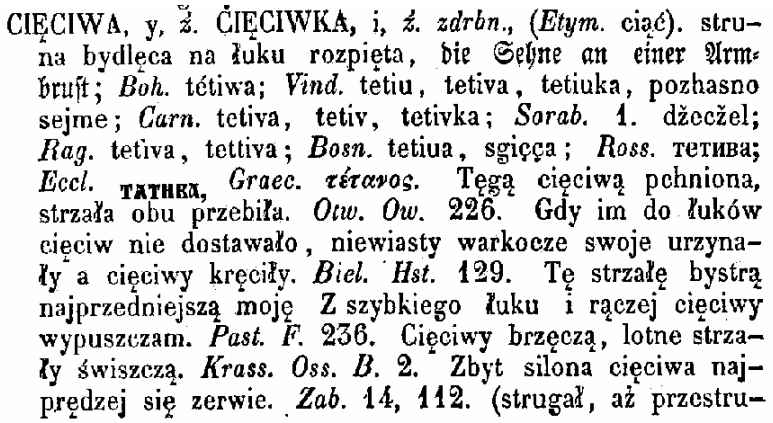
dictionary is a huge Słownik geograficzny Ziem Polskich… from the 19th c. It contains information on any geographical entity that at some time in history was in Poland; the electronic version was prepared for the Polish Genealogical Society in Chicago in DjVu technology by Rafał T. Prinke from Adam Mickiewicz University in Poznań. Unfortunately, the dictionary is for certain reasons not available, there are only individual copies of it. Another electronic version of this dictionary, in PDF images, is sold commercially (http://www.biblioteka.okay.pl/oferta.html). It is significant also that The Main Library of Warsaw University, which has a digitization project (http://www.buw.uw.edu.pl/zasoby/dygital.htm) agreed to scan some of the pre-nineteenth century dictionaries with Polish (Włodzimierz Gruszczyński, personal communication; unfortunately, they sell their products).

2. Dictionaries

The three historic dictionaries are:

- two editions of Słownik języka polskiego by Samuel Linde (1st: 1807–1814; 2nd: 1854–1860),
- Słownik języka polskiego, published in Vilnius (Polish Wilno), therefore called Słownik wileński (1861),
- Słownik języka polskiego, published in Warsaw, therefore called Słownik warszawski (1900–1927),

and a number of one volume dictionaries, for example those with so called foreign words, or the Polish version of the thesaurus by Roget by Zawiliński from 1926. In what follows I will describe briefly the dictionaries (for details see Piotrowski 1994, 2001). The dictionary by Linde is by far the most interesting of them, it is also the most dificult one to work on. Linde was a German, and he was not a native speaker of Polish. He produced a monumental dictionary, in 6 volumes, the first monolingual dictionary of Polish, and for a long time the first documentation dictionary as well. It had a great influence on dictionaries of other Slavonic languages, and perhaps on the dictionary by the Grimms. It has over 60,000 entries, with proper names (2,000). It includes not only Polish headwords, but each headword is translated into German, and Linde added also an enormous number (250,000) of equivalents from other Slavonic languages, and, when explaining etymologies, he uses also, for example, Greek and Hebrew words.

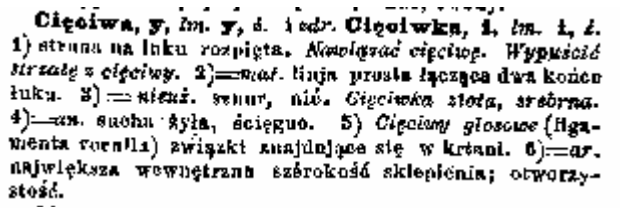


CIĘCIWA, y, ž. CIĘCIWKA, i, ź. *zdrbn.*, (*Etym.* ciąć). struna bydlęca na łuku rozpięta, die Sehne an einer Armbruſt; *Boh.* tétiwa; *Vind.* tetiu, tetiva, tetiuka, pozhasno sejme; *Carn.* tetiva, tetiv, tetivka; *Sorab.* 1. dżeczel; *Rag.* tetiva, tettiva; *Bosn.* tetiua, sgięça; *Ross.* тетива; *Eccl.* ТАТНВА, *Graec.* τέτανος. Tęgą cięciwą pchniona, strzała obu przebiła. *Otw. Ow.* 226. Gdy im do łuków cięciw nie dostawało, niewiasty warkocze swoje urzynały a cięciwy kręciły. *Biel. Hst.* 129. Tę strzałę bystrą najprzedniejszą moję Z szybkiego łuku i rączej cięciwy wypuszczam. *Past. F.* 236. Cięciwy brzęczą, lotne strzały świszczą. *Krass. Oss. B.* 2. Zbyt silona cięciwa najprędzej się zerwie. *Zab.* 14, 112. (strugał, aż przestru-

Figure 2 Linde's Dictionary

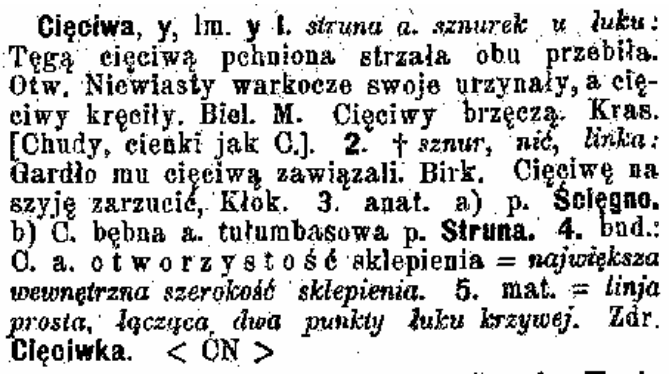
All words are printed in the proper script (this can be appreciated in Figure 2), therefore interpretation of this dictionary by any OCR application is an enormous challenge; on the other hand, it is very well known and it was reprinted twice, and was described in detail. The first edition, which some scholars say is superior to the second, is not so well known, that is why in the project both editions would be included

Słownik wileński is considered to be the first concise dictionary for common use. It has only two volumes, but, thanks to small print and compact definitions, it includes a huge number of headwords (110,000). It served as a model for many later dictionaries in the format of the entry. The print is very small, but quite clear and modern-looking, and the dictionary was re-printed once. The third dictionary, Słownik warszawski, has not been researched as well as the other two dictionaries. It is certainly the largest dictionary of Polish, but specialists disagree on basic things, for example the number of headwords, the highest given is 280,000. It is very rich in information, includes standard, dialectal, and colloquial words, which covers also obscene ones, expressions in (then) current use and obsolete ones, provides also synonimic expressions and collocations. It is still invaluable for its coverage of rare and very rare lexical items. Fortunately the dictionary uses predominantly characters in the Latin script, so it should not be very difficult to interpret. This is the most important dictionary that the project is to digitize.



Cięciwa, y, *lm.* **y**, *d.* i *zdr.* **Cięciwka**, i, *lm.* i, *d.* 1) struna na łuku rozpięta. *Nawijać cięciwę. Wypuścić strzałę z cięciwy.* 2)=*mat.* linja prosta łącząca dwa końce łuku. 3)=*ubiór.* sznur, nić. *Cięciwka złota, srebrna.* 4)=*an.* sucha żyła, ścięgno. 5) *Cięciwy głosowe* (Ligamenta vocalia) związki znajdujące się w krtani. 6)=*ar.* największa wewnętrzna szerokość sklepienia; otworzystość.

Figure 3 Słownik wileński



Cięciwa, y, lm. **y** i. *struna a. sznurek u łuku:* Tęgą cięciwą pchniona strzała obu przebiła. Otw. Niewiasty warkocze swoje urzynały, a cięciwy kręciły. Biel. M. Cięciwy brzęczą. Kras. [Chudy, cienki jak C.]. **2.** † *sznur, nić, linka:* Gardło mu cięciwą zawiązali. Birk. Cięciwę na szyję zarzucić. Kłok. **3.** anat. a) p. **Ścięgno.** b) C. bębna a. tułumbasowa p. **Struna.** **4.** bud.: C. a. otworzystość sklepienia = *największa wewnętrzna szerokość sklepienia.* **5.** mat. = *linja prosta, łącząca dwa punkty łuku krzywej.* Zdr. **Cięciwka.** < ON >

Figure 4 Słownik warszawski

The diachronic dictionaries are:

- Słownik staropolski (1953–2003)
- Słownik polszczyzny XVI wieku (1966–)
- Słownik staropolski, on which Polish scholars worked for a century, was finished, and has eleven volumes, however, because of the large amount of new material, the editors plan to add a supplement of several volumes. The characters are to a large

extent modernized, and do not diverge too much from the Latin script. Conversion of the dictionary to a digital form, and to text, will allow one to combine the text proper and the planned supplement, and, what is even more important, to link the interpretation of the meaning of the lexical items to the text in the corpus, as the compilers of the dictionary are busy now at creating a corpus of Old Polish (Twardzik and Górski 2003), therefore the dictionary and the corpus, when finished, will offer two complementary points of view on Polish from that period.

- Słownik polszczyzny XVI wieku has not been finished yet. It is the most ambitious lexicographic project in Poland, with a file index that has 8 million slips with citations. The dictionary will have about 70,000 entries, and the thirty-one published volumes include now about 50% of this material (they start entries with R now). It provides detailed information about semantics and syntax, but also statistics on the use of lexical items.

3. Technology: Scanning

In projects from more priviledged countries the texts were keyboarded rather than scanned (cf. the projects of the dictionaries by the Grimms, Johnson, or OED). In some projects two teams were even used, which independently keyboarded the text, which was later compared to reveal any errors. This method has numerous advantages, it is safe, and produces text that can be marked-up already at the data capture level. But it has several drawbacks as well: it is expensive, and calls for human resources which are far beyond the reach of our project.

After careful consideration, we decided to scan the dictionaries by contracting a specialized company, rather than to do it ourselves. There are several reasons. First, we would need a specialist library scanner, because the dictionaries are in thick volumes, and the shadow in the middle of a spread out volume will make interpretation by means of OCR far more difficult, if not impossible, what is more, the books should not be forcibly flattened, as this might damage them. Library scanners allow for non-intrusive scanning, and for elimination of the shadow (by hardware and software) as far as possible. However, they are very expensive, and the cost of the scanning services will not be much more expensive than the cost of one high-quality scanner, especially if the cost of maintenance, servicing and labour is to be taken into consideration. It is also important that the company, after working with some research libraries for other projects, has good relations with them, and is allowed to use the books from the holdings of those libraries. Some of the dictionaries are quite rare, and this aspect should not be underestimated.

The pages will be scanned as TIF images, 600 dpi, to preserve the details as accurately as possible. Then they will be converted into JPEG format. It will be necessary to convert them into even more compressed formats, as they will be made available at some Internet sites, and, with the high quantities of data to be downloaded, this aspect is certainly worth paying attention to. There will be two formats used: Adobe Systems' Portable Digital Format (PDF), and DjVu. DjVu is an image compression technology, developed in the 1990's at AT&T Labs, and later acquired by an American company, Lizardtech (http://www.lizardtech.com).

PDF, the Adobe Systems technology, is described by the producers as follows "Portable Document Format (PDF) is the de facto standard for the secure and reliable

distribution and exchange of electronic documents and forms around the world, with a ten-year track record. PDF is a universal file format that preserves the fonts, images, graphics, and layout of any source document, regardless of the application and platform used to create it. Adobe® PDF files are compact and complete, and can be shared, viewed, and printed by anyone with free Adobe Reader® software … You can convert any document to Adobe PDF using Adobe Acrobat® software products, enabling business, engineering, and creative professionals to create, distribute, and exchange secure and reliable Adobe PDF documents" (http://www.adobe.com/products/acrobat/adobepdf.html). The format has the advantage of being well known, and well supported, in that there are numerous free applications that can generate PDF files from various input (for example, the free OpenOffice package). However, for multi-page documents the files are not small enough, and it will take a long time to download them from the Internet.

DjVu allows the distribution on the Internet of very high resolution images of scanned documents, including images. It seems to be a superior format when the size of the files is considered, and download time is shorter. A DjVu file can theoretically be 1,000 times smaller than a TIFF file, and from 10 to 100 times smaller than a JPEG or a PDF file. In practice, a 3,5 MB TIFF image converted into DjVu (using an old free application, DjVu Solo) was compressed to 580 KB. One significant problem is that mass conversion of thousands of pages using the Windows platform is quite expensive, because there are only commercial products to actually make DjVu in batch processing, and these tend to be very expensive (the cost is thousands of dollars). There is a solution, however, in that for the Linux platform there are free packages that can be used for that purpose (http://djvu.sourceforge.net/). The resulting files are not as small, however, as those when the commercial products are used.

4. Technology and Analysis: OCR, encoding

The images of all of the dictionaries will be run through an OCR program. The choice of the OCR application has to be a compromise between efficiency and cost. Taking these two factors into account the Russian application Fine Reader 7.0 was chosen

Chrap, «, Im. y, m. i zdr. Chrapka, ł, 1) gniew ukryty, żądza tajemna szkodzenia. Mieć chrap na kogo. 2) = chętka zawladania czem, podstępem lub gwał-tem. Mieć wielki chrap na co. 3) == ten co chrapa, chrapa-ła. 4) ─ v. chrapy końskie, nozdrza u konia. 5) = prow. bagno, trzęsawisko. Strzelał kaczki dzikie na chrapach. Uwiązl w chrapach. 6) = v. nieuz. Cłtrapięc, ęci, Im. ecie, ź, Chrapięcia, y, Im. y, z, Chrapowina, y, Im.y z. niskie chropawe miejsce śród lasu, zarośla; krza-ki na miejscach mokrych. Nie bez trudności przebyłem te chrapy.

Figure 5 Result of OCR of Słownik Wileński

(from ABBYY). It offers high quality for its price, can be taught unknown characters, and, what is very important, it is language-aware and supports Polish, and can read Cyrillic without any special add-ons, as well as other "exotic" characters. The results of the OCR of a page from Słownik wileński can be appreciated in Fig. 5. below, which has not been post-edited at all. The print is very small, which is usually an obstacle for OCR applications, but otherwise the extract can be read without too much difficulty. In other words, the results are highly promising. There are no plans to mark-up the text in any way, the main thrust of the project is to make the images and the text available.

5. Dissemination

The dictionaries, as images, will be made as widely available as possible. They will be placed on DVDs (all of them) and CDs (individual dictionaries), which will be offered to the National Library in Warsaw, as well as to major research libraries in Poland; national libraries in the Slavonic-speaking countries will also be sent the disks. Free copying of the disks will be also encouraged. Hopefully, some of the Internet libraries will host the dictionaries (for example, Polska Bilbioteka Internetowa http://www.pbi.edu.pl/). Perhaps also the dictionaries as text will be made available "as is", as they will require extensive post-editing, or the interested individuals will receive the text to do post-editing. For example, Magdalena Majdak from Warsaw University is writing her PhD on Słownik warszawski and she is naturally interested to obtain the text of the dictionary to do her research.

References

1. Burch, Thomas, et al, *Ein "Hausbuch" für alle? - Das Deutsche Wörterbuch von Jacob und Wilhelm Grimm auf CD-ROM und im Internet,* In: *Jahrbuch für Computerphilologie 2,* Paderborn 2000, 11–34 [http://www.dwb.uni-trier.de/Beitrag_CP.htm]
2. Burch, Thomas, J. Fournier and K. Gärtner, *Mittelhochdeutsche Wörterbücher im Verbund,* CD-ROM und Begleitbuch. Stuttgart, 2002
3. Christmann, Ruth, *Books into Bytes: Jacob and Wilhelm Grimm's Deutsches Wörterbuch on CD-ROM and on the Internet*, Literary and Linguistic Computing 1 (2001), 121–133.
4. Christmann, Ruth, and Thomas Schares, *Towards the User: The Digital Edition of the Deutsche Wörterbuch by Jacob and Wilhelm Grimm*, Literary and Linguistic Computing 18:1 (2003).
5. Králík, J. and P. Šmídová, *Elektronická podoba slovníku Spisovného jazyka českého*, Slovo a slovesnost 61:4 (2000), 318–320
6. McDermott, Anne (ed.), *Samuel Johnson's Dictionary on CD-ROM,* Cambridge Univ. Press, 1996.
7. McDermott, Anne, *Creating an Electronic Edition of Johnson's Dictionary: Developments of Solutions to Some Problems,* in: Thomas Burch, Johannes Fournier, Kurt Gärtner, and Andrea Rapp (eds.), *Standards und Methoden der Volltextdigitalisierung,* Mainz: Akademie der Wissenschaften und der Literatur, 2003, pp. 153–60
8. Piotrowski, T., *Z zagadnień leksykografii,* Warszawa: Wydawnictwo Naukowe PWN, 1994
9. Piotrowski, T., *Zrozumieć leksykografię,* Warszawa: Wydawnictwo Naukowe PWN, 2001
10. Ross, S., M. Donnelly and M. Dobreva, *New Technologies for the Cultural and Scientific Heritage Sector* (DigiCULT, Technology Watch Report 1), Salzburg: European Commission, 2003
11. Twardzik, W. B. and R. L. Górski, *Korpus staropolski Instytutu Języka Polskiego PAN w Krakowie,* W: S Gajda (Red. nauk.), *Językoznawstwo polskie. Stan i perspektywy,* Warszawa, Opole: PAN, Uniwersytet Opolski, 2003, ss. 143–154.

12. Unsworth, John, K. O'Brien O'Keeffe and L. Burnard (eds.), *Electronic Textual Editing*, the Modern Language Association and the TEI Consortium, 2004 (w druku)
13. Weiner, Edmund, *The New Oxford English Dictionary: Progress and Prospects,* In: R. W. Bailey (ed.), *Dictionaries of English. Prospects for the Record of our Language*, Ann Arbor: University of Michigan Press, 1987, pp. 30–48.

tadpiotr@plusnet.pl