Digitalizing dictionaries of Polish*

Janusz S. Bień
Formal Linguistics Department, University of Warsaw
jsbien@uw.edu.pl

October 5, 2008
March 2, 2009

Abstract

The author has been actively involved in digitalization of several dictionaries, including the 17th century Knapski's dictionary, the 18th century dictionary of Troc, the 20th century so called „Warsaw dictionary" and some volumes of the work-in-progress dictionary of Polish language of the 16th century. All of the available dictionaries are in the DjVu format and the electronic versions have been created mostly with free DjVuLibre software tools. The advantages of the format from a user point of view are discussed.

Keywords: dictionaries, Polish, digitalization, DjVu, DjVuLibre

1 Principal dictionaries of Polish

The comprehensive lexicographic description of Polish is to be provided by the series of great dictionaries:

- *Słownik staropolski* (The Old Polish Dictionary), which covers early Polish language up to year 1500. Published in 1953–2003 (ISBN 83-04-00472-0).
- *Słownik polszczyzny XVI wieku* (Dictionary of the 16th century Polish). The work started in 1949 and is still in progress.
- *Słownik języka polskiego XVI i pierwszej połowy XVIII wieku* (the Dictionary of Polish language of 17th and the first half of the 18th Century), intended to fill in the gap between the dictionary of the 16th century and the dictionary mentioned below. Work on it started in 1954 and is still in an early stage.
- *Słownik języka polskiego PAN* (Polish Academy of Sciences Dictionary of Polish), commonly known as Doroszewski's dictionary after the name of

*Please consult also slides available at http://bc.klf.uw.edu.pl/71.

its chief editor. Published in 1958–1969 (ISBN 83-01-11876-8), intended to cover contemporary Polish language of that time, starting with the second half of the XVIII century (to include the language of some important literary works).

- *Wielki słownik języka polskiego* (Great dictionary of Polish). The dictionary is in the preliminary stage of development and its purpose is to describe contemporary Polish.

On the other hand, one of the most demanded Polish dictionaries is not a dictionary of Polish, but *Słownik geograficzny Królestwa Polskiego i innych krajów słowiańskich* (The Geographical Dictionary of the Polish Kingdom and other Slavic Countries), a gazetteer in 15 volumes of almost 1000 pages each published in 1880-1914, extremely useful for genealogical research. The gazetteer covers Poland in its borders before the partitions between Russia, Germany and Austria, but due to the censorship it was impossible to state this explicitly in the title.

1.1 The Old Polish Dictionary

As far as I know, there are not yet any specific plans to digitalize the dictionary.

1.2 Dictionary of the 16th century Polish

Słownik polszczyzny XVI wieku has been in preparation since 1949, for rather historical reasons in the Institute of Literary Research of the Polish Academy of Sciences in its Toruń and Wrocław annexes (the annex in Kraków no longer exists). 32 volumes appeared in print in years 1966-2004 (ISBN 83-04-00477-1), the last one covers the entries from Przemijać to Przodujący. The head of the team, Prof. Franciszek Pepłowski, was fully aware that the computer files used for typesetting should be preserved (perhaps it was the influence of the late Prof. Jerzy Woronczak, one of the members of the editorial committee and the pioneer of the use of computers in linguistics and lexicography), but it became practically feasible rather late, namely since volume 23 was printed in 1995. It was made possible by the fact that since then the dictionary was typeset locally by the members of the editorial team using commercial typesetting programs.

Although the meeting in 2001 with Prof. Pepłowski seemed to be the start of a very fruitful cooperation aiming at preparing both printed and electronic version of the dictionary at the same time (Piotrowski, Szafran 2005, Szafran 2007), the idea has been rejected by Prof. Pepłowski's authorities without providing any justification. The recent retirement of Prof. Pepłowski made the situation only worse. The position of the head of the Institute, Prof. Elżbieta Sarnowska-Temeriusz, expressed in a letter of 21.07.2008 to the chairman of the Social Sciences Division of Polish Academy of Sciences (the letter is the response to my intervention), is that all the new digitally-born volumes will be made available on the Internet just by scanning the printed copy. As a consequence, volume 32, published in 2004 and converted directly from printer's

PDF files to DjVu using the recently released pdf2djvu program (Wilk 2008:33) cannot be made available publicly, although it has been ready since May 2008.

In 2003 Prof. Tadeusz Piotrowski unsuccesfully applied for a grant to scan, among others, *Słownik polszczyzny XVI wieku*. In 2004 he talked about it on a conference, and his talk appeared in print (Piotrowski 2005). It was noticed by Peter Krolikowski, known for his work on Brześć's Bible (`http://www.biblia.net.pl/bb/brzeska.htm`), a translation of the Bible into Polish published by Calvinists in 1563. He contacted Piotrowski informing him that he has scanned for his own use the whole first volume of the dictionary and is willing to make the scans freely available. I in turn asked the authorities of the Institute of Literary Research in my letter of 10 October 2005 for a permission to publish them on the Internet.

For the present purpose it is sufficient to say that, after some struggles, in January 2006 I have managed to get the permission, although it were the better quality scans made by Krzysztof Opaliński (an editorial team member) which has been finally converted to DjVu format, augmented by me with various additional information and made available on the server of the Computer Science Institute of the University of Warsaw. A large format map included in this volume has been scanned courtesy of University of Warsaw main library. This work inspired the head of Kujawsko-Pomorska Digital Library to scan all the volumes of the dictionary. In particular the first volume was scanned again, but with much lower quality ... (looks like these scans have been replaced in June 2008 with slightly better ones). At the moment of writing (October 5, 2008), scans of the first 15 volumes are available (all in DjVu format)[1].

I maintain a Web page containing links to the digital versions of the dictionary (`http://www.mimuw.edu.pl/polszczyzna/fSpXVIw/`).

1.3 The Dictionary of Polish language of 17th and the first half of the 18th Century

Słownik języka polskiego XVII i pierwszej połowy XVIII wieku has a much more optimistic history. The first volume was printed in the traditional form, several fascicles appeared between 1977 and 2001 (ISBN 83-85579-94-X) closely following *Słownik polszczyzny XVI wieku* in the layout and structure. However, after the retirement of the former head of the unit it was decided to use Internet as the primary publication medium. This wise, although difficult, decision was taken by Prof. Włodzimierz Gruszczyński, the new head of the team since 1 January 2003. The software has been developed by Mateusz Żółtak. As a result, the current snapshot of the dictionary is always available at `http://xvii_wiek.ijp-pan.no-ip.org/pan_klient/index.php`.

[1] The remaining 17 volumes (including those digitally-born) has been scanned by December 3rd, 2008.

1.4 PAS Dictionary of Polish

Słownik języka polskiego PAN, prepared under the auspices of the Polish Academy of Sciences and originally published in 1958–1969, has been digitalized by the commercial PWN Scientific Publishers as the side effect of preparing its reprint published in 1996-1997. At the time I had an additional part-time job there in the department of Polish language dictionaries and proposed to publish also the scans of the dictionary. The idea was given a concrete shape by Mariusz Olko, who already cooperated with the publisher on typesetting and digitalization. The full story is to be told on some other occasion, the final outcome was that the CD of the so called electronic reprint was included as a free bonus with the printed dictionary and was also sold separately (ISBN 83-01-12321-4). The work was done under my supervision but under heavy time and financial constrains. In consequence, my attitude to the result is ambivalent; in some respects the CD can serve as an example of optimal solutions, but it has also some substantial flaws causes by the above-mentioned constrains.

1.5 Great dictionary of Polish

Wielki słownik języka polskiego is mentioned here only because by design its primary form will be an electronic version freely available on the Internet.

1.6 The Geographical Dictionary of the Polish Kingdom and other Slavic Countries

As for *Słownik geograficzny Królestwa Polskiego i innych krajów słowiańskich*, it has been digitalized independently in several ways, so it can be used also to conveniently compare different approaches to the problem. The first electronic version freely available on the Internet was prepared by me in DjVu format and has been available at `http://www.mimuw.edu.pl/polszczyzna/SGKPi/` since 2005.

1.7 Other important dictionaries

By other important dictionaries we mean here the dictionaries referred to in the principal dictionaries described above.

Słownik języka polskiego PAN regularily mentions whether the entry appears in Linde's słownik, „słownik warszawski" („Warsaw dictionary") and „słownik wileński" („Wilno's dictionary").

Słownik języka polskiego by Samuel Bogumił Linde has been published in 6 volumes in 1807-1814. The scans of this first edition are now available in Kujawsko-Pomorska Digital Library: `http://kpbc.umk.pl/publication/8173`.

Earlier an incomplete second edition of the dictionary has been made available in the highly controversial Polish Internet Library (`http://www.pbi.edu.`

pl)[2]. Since 2007 an incomplete first edition of Linde's dictionary is available also in Google Book Search (it was pointed to me by Prof. Tadeusz Piotrowski). The 5th volume is at http://books.google.com/books?id=kc4GAAAAQAAJ, other volumes has to be accessed as 'Other editions', which is evidently misleading.

„Słownik wileński" is a dictionary published in Wilno in 1861. The initiative to digitalize it has been put forward in 2007 by Małgorzata B. Majewska. Some preliminary work has been done in the form of e-SWil project (http://swil.zozlak.org/) using the software written by Mateusz Żółtak. The members of the project were students of the Institute of Polish Language of the University of Warsaw[3].

„Słownik warszawski", more precisely *Słownik języka polskiego* by Jan Karłowicz, Adam Kryński and Władysław Niedźwiecki, has been published in Warsaw in years 1900–1927. It has been scanned by the library of the University of Warsaw and converted by me to the DjVu format. The results are now available at the University of Warsaw Digital Library. Unfortunately, the library (contrary e.g. to Kujawsko-Pomorska Digital Library) does not provide single links to multivolume works. The first volume can be read at http://ebuw.uw.edu.pl/publication/255, to get to the other volumes the reader has to follow the „structure" (Struktura) link.

Słownik polszczyzny XVI i pierwszej połowy XVIII wieku refers additionally also to Knapski's and Troc's dictionaries. Both of them have been scanned by the library of the University of Warsaw and converted by me to the DjVu format.

The first part (Polish-Latin-Greek) of *Thesaurus Polonolatinograecus seu Promptuarium linguae Latinae et Graecae* by Grzegorz Knapski (alias Knapiusz, Knap) has been published in 1621[4], the second part (Latin-Polish) in 1626, the third part (*Adagia continens*) in 1632[5]. The second edition of the first and second parts has been published in 1643 - 1644 and is are available at http://www.mimuw.edu.pl/polszczyzna/Knapski/Knapski_DjVu/. This version incorporated an index prepared in 2004 by Marek Kunicki-Goldfinger and Aldona Przyborowska-Szulc. They can be also purchased on CD from the library of the University of Warsaw (ISBN 978-83-924821-1-6). The CD contains additionally a book about Knapski's dictionary by Jadwiga Puzynina, which recently became available also at http://ebuw.uw.edu.pl/publication/443.

Nowy dykcyonarz ... by Michał Abraham Troc (alias Trotz) with the parallel title *Nouveau dictionnaire polonois, allemand et françois enrichi de proverbes*

[2]The second edition is missing the first and the last volume. A recent check revealed that the incomplete first edition (missing the last volume) is also available there. Probably for the bureaucrats, who commissioned in 2002 the creation of the the library and spent a shocking amount of money on it, only the total number of digitalized items was important.

[3]This work will be supported by a grant *Edycja elektroniczna Słownika wileńskiego* (number N N104 175236) of the Ministry of Science and Higher Education awarded in February 2009.

[4]Since 30.01.2009 available in Digital Library of Wielkopolska courtesy of Biblioteka Kórnicka (http://www.wbc.poznan.pl/publication/92345).

[5]Since 12.02.2009 available in Digital Library of Wielkopolska courtesy of Biblioteka Kórnicka (http://www.wbc.poznan.pl/publication/93346).

les plus usitez, de remarques de grammaire, de termes de medecine, de botanique, de matematique, de fortification, de marine, de chasse et des autres arts appeared in 1764 in Leipzig (Lipsk). It is described as volume III, as the author considered it to be the sequel to the *Nouveau dictionnaire Francis, allemand et polonais* published in two volumes in 1744 and 1747. *Nowy dykcyonarz* is available at `http://ebuw.uw.edu.pl/publication/393`.

2 DjVu technology and DjVuLibre

2.1 History

The DjVu technology, described by its authors as *an image compression technique, a document format, and a software platform for delivering documents images over the Internet* (Le Cun 2001:2), was originally developed by Yann Le Cun, Léon Bottou, Patrick Haffner, and Paul G. Howard at AT&T Laboratories in 1996. The implementation of the DjVu technology available on the GNU GPL licence (`http://djvu.sourceforge.net/lti-licensing.html`) is called DjVuLibre.

The DjVu technology has several aspects. First of all, it provides very efficient algorithms for image compression; best of them are still available only in the form of commercial and quite expensive products. Secondly, it provides an efficient way to transfer the compressed images over the Internet, even on relatively slow lines. Moreover, it provides also an efficient way to display the image on the end-user's computer, using such tricks as progressive decoding (only this part of the image is decompressed which is to be displayed), downloading next page in the background etc.

As dictionaries are large and its pages are not accessed in a sequential way, it is of crucial importance that DjVu allows to store every page in a separate file and download only the pages which are really needed.

2.2 Applications

One of the first applications demonstrating the advantages of DjVu and still one of the best examples of a digitalized dictionary is *The Century Dictionary* available since 2001, now at `http://www.global-language.com/CENTURY/`. The recently redesigned dictionary site is now missing the description of the project, fortunately it is still available at `http://web.archive.org/web/20010713134356/216.156.253.178/CENTURY/why.php`. To make a long story short, it was prepared almost single-handedly by Jeffery A. Triggs, at that time a member of the DjVu development group at AT&T Labs. Let's quote his statement: «we created *The Century Dictionary Online* because it is *free*, it is *big*, and it is *beautiful*. I should add finally, a couple of more reasons: married to DjVu technology, it is *innovative*, ... ».

The reader is invited to compare DjVu technology to other formats by consulting various versions of *Słownik geograficzny*: the entry Wilno in National

Digital Library at http://www.polona.pl/Content/6259/15760_Wilno.html, the same entry in the version available in the International Centre for Mathematical and Computational Modelling (http://dir.icm.edu.pl/pl/Slownik_geograficzny/), in Małopolska Digital Library (http://mbc.malopolska.pl/publication/113) and in the DjVuLibre edition (http://www.mimuw.edu.pl/polszczyzna/SGKPi/). In the last two experiments you can be guided by the tutorial produced by the Toronto Ukrainian Genealogy Group (http://www.torugg.org/Publications/How_to_use_Polish_Gazeteer_Online.pdf).

It may be also interesting to compare Linde's dictionary in Kujawsko-Pomorska Digital Library, in Polish Internet Library and in Google Book Search.

2.3 Viewers

From a user point of view it is the DjVu viewer which is important. There exist several of them, both commercial and free, for various platforms, palmtops and cellular phones included. All the viewers profit from the DjVu design features allowing the viewer to simulate the operations on a paper document in comparable time, as illustrated by the table 4 in (Le Cun 2001:6):

Action	Real-word equivalent	Acceptable delay
Zooming/Panning	Moving the eyes	Immediate
Next/Previous Page	Turning a page	< 1 second
Random Page access	Finding a page	< 3 seconds

I would like to draw the reader's attention to a feature of a relatively new multiplatform viewer djview4 (cf. http://djvu.sourceforge.net/djview4.html), which is available for Debian GNU/Linux, MS Windows and MacOS. It can be used to browse both local and remote documents.

From the very beginning, the DjVu viewers allowed a specified fragment of the text to be highlighted. For example, the address http://www.mimuw.edu.pl/polszczyzna/SGKP/SG06.djvu?djvuopts=&page=tom06-337.djvu&highlight=2003,1827,1679,210 points to a specific fragment of *Słownik geograficzny*. The main part of the address describes the primary document file, which in this case is just an index to the files containing individual pages of the 6th volume of the dictionary. The parameter page describes the page using its name which happens to coincide with the name of the file containing it. The highlight parameters specifies coordinates in pixels of the rectangle to be highlighted. Now the practical question is: how to calculate the coordinates for a specific rectangle?

Upon my request, a student of mine Jakub Wilk extended the viewer with the facility to mark a fragment with a mouse and than put the appropriate address into the clipboard. He submitted the patch on 9.02.2008. By 29.02.2008 this feature has been reimplemented in a more efficient way by Léon Bottou and is now available in the standard distribution of the viewer.

I think this feature is quite important for academic research, as it allows to quote a specific fragment of digitalized work when including its image is technically difficult or not desirable.

3 Conclusion

Dictionaries are a special kind of publications used in a specific way. Although DjVu technology itself is not sufficient for a fully functional digitalization of dictionaries, it is extremely useful as the first step in this direction.

Acknowledgments

The paper has been prepared using LaTeX authoring system. Joanna Biliń-ska adapted it to the editors requirements. It was proofread by Radosław Moszczyński.

References

Le Cun Y, Bottou L., Erofeev A., Haffner P., Riemers B. (2001), "DjVu document browsing with on-demand loading and rendering of image components" in *Internet Imaging*, San Jose, January 2001a. `http://leon.bottou.org/papers/lecun-2001`.

Piotrowski T. (2005), "Digitization of Polish historic(al) dictionaries". *Review of the National Center for Digitization* Issue 6 (2005), pp 95-102. `http://www.ncd.matf.bg.ac.yu/casopis/06/Piotrowski/Piotrowski.pdf`.

Piotrowski T., Szafran K. (2005). "The dictionary of Polish of the 16th century and the computer: from paper to a (structured) file" in F.Kiefer, G.Kiss, J.Pajzs (eds.), *Papers in Computational Lexicography Complex 2005*: Proceedings of the 8th International Conference on Computational Lexicography, Complex 2005, June 17-18, Budapest, Hungary, 171-179.

Wilk J. (2008). Rozbudowa pakietu oprogramowania DjVuLibre. M.Sc. thesis, Computer Science Institute of the University of Warsaw, 2008. `http://jw209508.hopto.org/papers/thesis/`.

Szafran K. *Analiza i formalny opis struktury „Słownika polszczyzny XVI wieku"*. Wydawnictwa UW 2007. `http://ebuw.uw.edu.pl/publication/253`.

The paper metadata in BibTEX format:

```
@incollection{bc71,
          author = {Janusz S. Bie\'n},
       booktitle = {Methods of Lexical Analysis:
                    Theoretical assumption and practical applications},
          editor = {Krzysztof Bogacki  and Joanna Cholewa and Agata Rozumko},
         address = {Bia\l{}ystok},
           title = {Digitalizing dictionaries of Polish},
       publisher = {Wydawnictwo Uniwersytetu w Bia\l{}ymstoku},
           pages = {37--45},
            year = {2009},
             url = {http://bc.klf.uw.edu.pl/71/}
}
```