Polish digital libraries as a text corpus

Filip Graliński

Adam Mickiewicz University
Faculty of Mathematics and Computer Science
Umultowska 87, 61-614 Poznań, Poland
filipg@amu.edu.pl

Abstract

A large (71 GB), diachronic Polish text corpus extracted from 87 digital libraries is described in this paper. The corpus is of interest to linguists, historians, sociologists as well as to NLP practitioners. The search engine based on the corpus is compared to Google Ngram Viewer. The sources of noise in the corpus (metadata and OCR errors) are described and assessed. Finally, a simple experiment to assess the impact of the noise on search results is discussed.

Keywords: digital libraries, text corpora, OCR

1. Introduction

According to Eric Schmidt[1] , every two days humanity creates as much information as it did up to 2003 (estimated at 5 exabytes (Lyman and Varian, 2003)). With the constant influx of new web pages, e-mails, blog posts, instant messages and tweets it might be easy to forget about the *very first* exabytes of humanity. It might be easy to forget, unless one is a historian, a linguist, a philologist or a genealogist – then having at hand such "paleodata" would be a fulfillment of a long-awaited dream.

And this dream is coming true, as more and more efforts are made to preserve cultural heritage all over the world and more and more print material – old newspapers, books, documents, posters, photographies and even school certificates[2] or train tickets[3] – is being digitised and made available online.

In particular, vast amounts of information are becoming available due to continuous and systematic digitisation initiatives to create *digital libraries* (collections of content that are both digitised and organised). Do Polish digital libraries keep up with the world's best? Yes, they do, actually, successful efforts made them stand out from the global pack as far as the quantity, the quality and the availability of digitised content is concerned. In this paper, we are going to discuss the content available within Polish digital libraries both in terms of metadata and full text.[4] Our aim is to look at the textual content digitised within digital libraries as a text corpus and present challenges they pose and *offer* for the natural language processing community.

With a diachronic corpus in hand (and, as we shall see, the data from Polish digital libraries represent a massive diachronic text corpus), it is natural for a linguist or a historian to look for the earliest mention of a word or phrase. This could be done for purely linguistic reasons, for example for antedating dictionary entries in lexicography, i.e. collecting citations predating those specified in a given dictionary (Wierzchoń, 2010; Podhajecka, 2010), or with a broader – historical, sociological, philological etc. – perspective in mind ("when exactly did a phenomenon emerge?", "when a notion was referred to for the first time?"). Technical problems soon become apparent once one undertakes such an enterprise: the quality of optical character recognition varies substantially (in general, the older the document, the worse the quality of the recognised text), noisy metadata (e.g. timestamps) lead to spurious („anachronic") search results. In this paper, we are going to discuss such issues from an NLP perspective.

This paper draws upon the seminal work of Piotr Wierzchoń (Wierzchoń, 2009; Wierzchoń, 2010), we address the issue from a more technical point of view here.

2. Polish digital libraries

The Web site of Polish Digital Libraries Federation[5] lists 92 digital libraries (excluding ones marked as „moved" or „in progress").[6] This is an impressive number and even more impressive is the number of all items available from Polish digital libraries: 1,310,209 (that is the number of metadata records we managed to collect from 87 libraries).[7] The distribution is not uniform, the two largest digital libraries (Jagiellonian Digital Library[8] and Digital Library of Wielkopolska[9]) account for 32% of all

[1] http://techcrunch.com/2010/08/04/schmidt-data/

[2] See e.g. http://www.sbc.org.pl/Content/51267

[3] See e.g. http://dlibra.karta.org.pl/catl/Content/13495

[4] Current studies of digital libraries in Poland seem to focus on the metadata and the selection of material to be digitised rather than on what is already available in terms of full-text search. The discussion of full-text search is rather scarce, see e.g. the presentation by Arkadiusz Pulikowski available at http://www.ptin.org.pl/konferencje/10forum/repozytorium/Pulikowski.pdf

[5] http://fbc.pionier.net.pl/owoc/list-libs

[6] All the data presented here were gathered in June 2013.

[7] We were not able to collect metadata from some libraries either because a library does not provide easily accessible metadata or due to technical problems. Also, not all records were collected for some libraries, again, because of technical issues.

[8] http://jbc.bj.uj.edu.pl

[9] http://www.wbc.poznan.pl

Figure 1: Number of metadata records collected from digital libraries (June 2013)

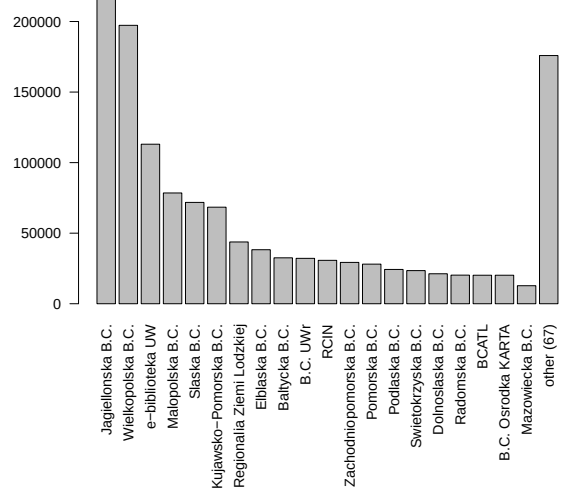
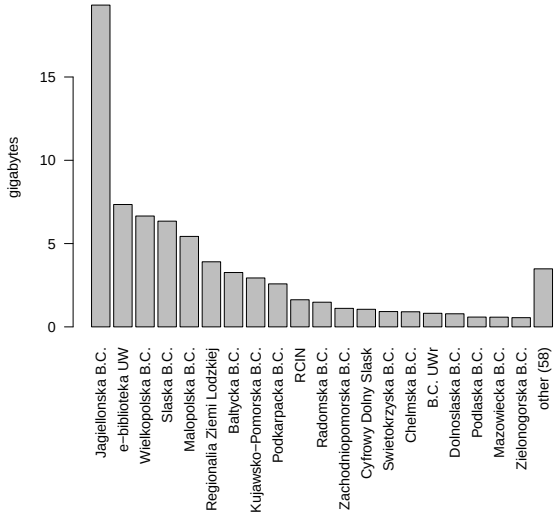


Figure 2: Size of plain text



the metadata records, whereas as many as 43 libraries provide no more than 2000 records, see Figure 1.

The majority of material in Polish digital libraries is available in DjVu or PDF formats, rarely in purely graphical formats, such as JPEG. Most digital libraries use dLibra framework[10] for storing and serving their collections.

3. Textual content available in Polish digital libraries

Although some items available in Polish digital libraries are non-textual (e.g. photographs or pictures), most of them are of a textual nature (newspapers, magazines, books) – after filtering out the metadata clearly referring to non-textual contents, 1,259,882 (96%) records were left.[11] As we were interested only in the Polish language, we limited ourselves to those items that were marked as Polish in the metadata or that were missing language information (presumably most of them are Polish as well). This way, we obtained 1,069,708 records.

Over a million metadata records is an impressive number, but we are interested in *full texts* rather than just titles, descriptions, author names, etc. Fortunately, for some (albeit not all) publications, full text is available embedded in DjVu or PDF files as a plain text layer. And, what is even more advantageous, most digital libraries provide hyperlinks for search engines leading to such plain-text versions of their publications, for instance, a publication available in Małopolska Digital Library at `http://mbc.malopolska.pl/dlibra/doccontent?id=13815&dirids=1` (in DjVu format) is also available as plain text (generated with OCR) at `http://mbc.malopolska.pl/dlibra/plain-content?id=13815`.

The total size of plain text corpora collected was over 71 GB. The corpus size broken down by libraries are presented in Figure 2.

How *diachronic* is the corpus? Figure 3 shows the distribution of plain-text material extracted from digital libraries over 5-year periods (publication date as specified in metadata was used, we were unable to determine publication year for about 10% of items). As it can be clearly seen, it is the interwar period that is best represented in Polish digital libraries. Nevertheless, there is a substantial amount of text for all the years since the mid-nineteenth century. The following factors were likely to come into play in determining the shape of the distribution:

- world wars (objective decrease in the number of publications),
- legal issues (in particular, 70-year copyright term for newspapers),
- policies of the different digitial libraries (they vary from library to library, for instance, Digital Library of Wielkopolska seems to systematically digitise regional newspapers from the earliest years onwards – consequently regional press of the second half of 20th is still scarce there, whereas Małopolska Digital Library digitises regional newspapers in a less chronological manner).

The text corpus extracted from digital libraries can be compared to the corpus used in Google Books Ngram Viewer[12] (a graphing tool which displays the changes in frequency of words or phrases over the selected years). The

[10] Developped at Poznań Supercomputing and Networking Center, see `http://dlibra.psnc.pl/`

[11] Due to noise and inconsistency in metadata some non-textual item might haved passed through the filtre.

[12] http://books.google.com/ngrams/

Figure 3: Size of plain text across five-year periods

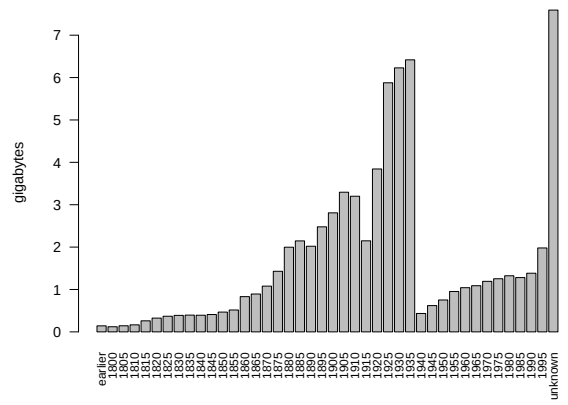


Figure 4: The number of occurrences of words *telewizor* (*TV set*) and *komputer* (*computer*) in years 1920-2000

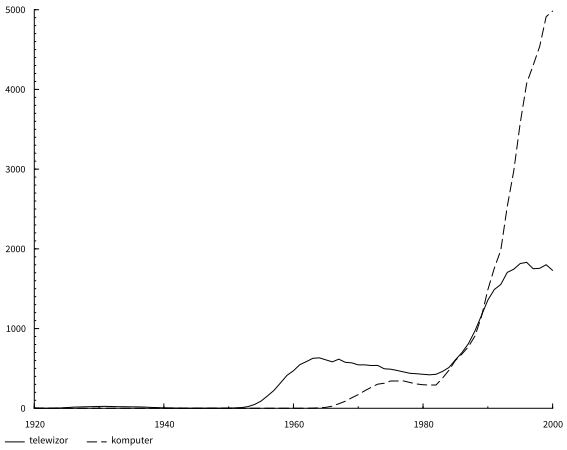


Figure 5: The number of occurrences of words *NRF* and *RFN* in years 1950-2000

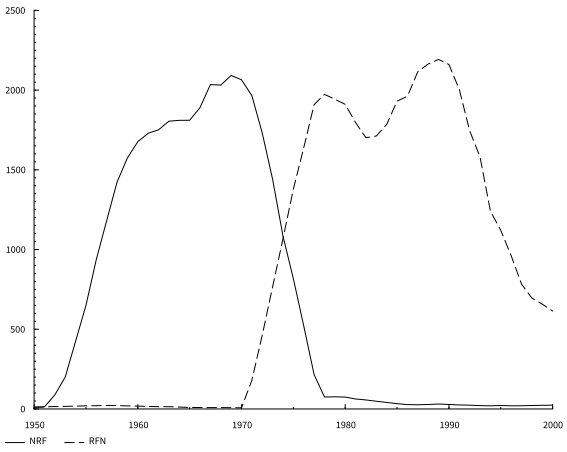
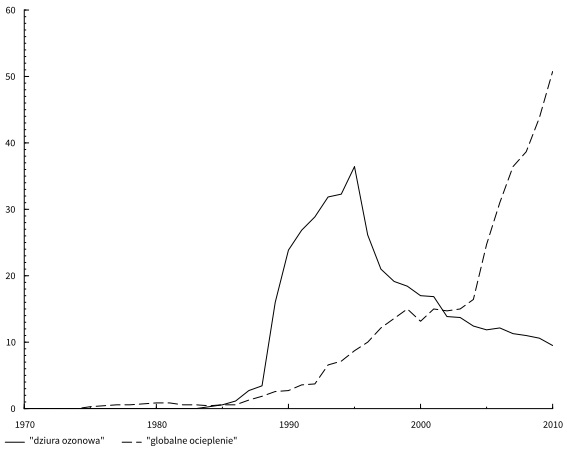


Figure 6: The number of occurrences of phrases *dziura ozonowa* (*ozone depletion*) and *globalne ocieplenie* (*global warming*) in years 1950-2000



Google Ngram Viewer corpus was composed of over 5 million digitised books – „ 4% of all books ever published" (Michel et al, 2011).

As Polish was not included among the seven Ngram Viewer languages (English, French, Spanish, German, Chinese, Russian and Hebrew), the corpus extracted from Polish digital libraries is probably the only way to get comparable results for the Polish language. For instance, see figures 4, 5 and 6 for graphs similar to NgramViewer frequency graphs (a smoothing of 3 was used for these graphs[13], all the inflected forms were taken into account).

Figure 4 shows how the words referring to two modern inventions have occurred in the corpus – it was in mid-1990s when the "computer" surpassed clearly the "TV set". Note that the word *telewizor* was used as early as in 1920s.

Figure 5 compares and contrasts two Polish abbreviations for West Germany: *NRF* (*Niemiecka Republika Federalna*) was used first, then (in 1970) the communist authorities changed it arbitrarily to *RFN* (*Republika Federalna Niemiec*). The switch is clearly visible in the graph.

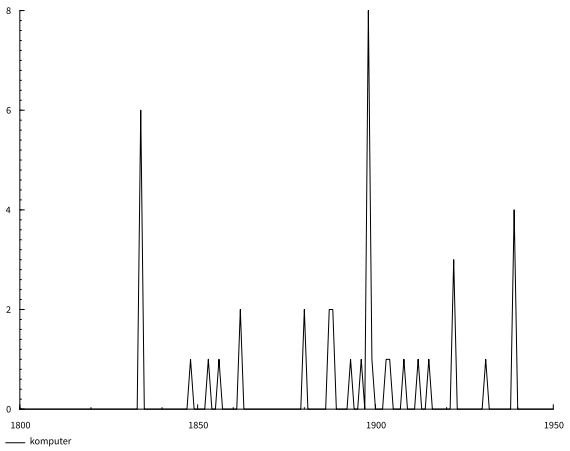
Finally, figure 6 visualizes media attention to two closely related (yet distinct) phenomena, global „threats": *dziura ozonowa* (*ozone depletion*) and *globalne ocieplenie* (*global warming*). *Dziura ozonowa* peaked in mid-1990, when it was a very popular subject with journalists, now it is mostly forgotten by the general public.

4. Sources the noise in the corpus

If, while antedating a word or drawing a graph of word frequencies, one were to take the results of full-text search at face value, one would sometimes come to utterly absurd conclusions. For instance, in pre-1950 texts one would find

[13] I.e. the data shown for a given year is an average of the count for the year plus 3 values on either sides.

Figure 7: The number of occurrences of word *komputer* (*computer*) in years 1800-1950 (no smoothing)



spurious quotations for the word *komputer* (*computer*)[14], see figure 7.

There are two major sources of noise in the text corpus extracted from the digital libraries:

- metadata errors,
- OCR errors in the full texts.

Metadata errors

It is the wrongly specified date of creation/publication that is the most troublesome type of metadata errors (at least for antedating), as "anachronic" items are very likely to pollute search results when they are ordered by date ascending. For instance, one spurious pre-1950 quotation for *komputer* originated from an article from February 1988 issue of "Tygodnik Radomski" whose Dublin Core *Date* element was entered as "1888-03-02" (instead of "1988-03-02"). Another example: 4 issues of "Grom Mazowsza" weekly were dated 1391 instead of 1931.

Even though such trivial errors are not frequent (for instance, there were only 11 items dated before 1700 and described as periodicals, a clear indication of error, as the first Polish newspapers were published in 18th century), they tend to come again and again in search results.

One simple method to filter out metadata errors is to make use of some ad hoc heuristics, for instance:

- simple conditions for too early (or late) years, e.g.: before 1700 for a periodical, before 1300 for any item in Polish, after the current year for any item,
- confronting the creation date with the year which is sometimes specified in the Dublin Core *Title* element, e.g. "Tygodnik Radomski, 1988, R. 7, nr 9" was specified as the *Title* element for the misdated issue of "Tygodnik Radomski" mentioned above.

[14]Contrary to English, in Polish *komputer* has only been used for an apparatus, never for a person who makes calculations.

A more advanced technique would be to apply a system for automatically determining the publication dates of a document from its contents. See (Garcia-Fernandez et al, 2011) for a description of such a system (for French) based both on supervised and unsupervised learning, and using external resources (e.g. Wikipedia biographical entries) and etymological knowledge. A corpus text extracted from digital libraries is in itself an ideal resource for developing, training and testing such systems.

Note that even in absence of trivial errors, it is a challenging and non-trivial problem to extract and normalise creation dates specified in the metadata – dates are given in various formats ("1951-02-07", , "1951.02.07", "7 II 1951 r.", etc.) there is hardly any consistency even within a single digital library.

Another problem is that a very long year span is sometimes specified for metadata. For instance, most issues of "Stolica" weekly magazine are dated just "1946-1989". If one were to take the lower end of such a span, many words would be incorrectly antedated (e.g. the word *komputer*). Obviously, it is more "safe" to assume the upper end when searching for the earliest occurrence of a word.

OCR errors

The word *komputer* is "attested" in the text corpus extracted from digital libraries as early as 1834:

> Jutro a Bogaskiego przy ulicy Długiej pod S50, w domu dawniej B aide go, a teraz Sammt ŚNIADANIE: Zupa pomidorowa, Krupnik gospoó ski, Pieczeń wołowa z rolna 5 mssłetn piecza; wetn, Zrazy z kaszą grzybowy, Wątróbka ciel ¿ szpikowana Z rożna, Mostki cielęce faszerowp Potrawa z pulard % kalafjorami, Legnmina z pie śliwkami, Kapłonki młode z komputera, etc.

Of course, this is a false positive caused by an OCR error, the word that is used in the text is *komputem* rather than *komputerem*, see figure 8[15] *komputem* is the instrumental singular of *komput*, a dialectal form of *kompot* (*compote*).[16]

The word *komputem* recognised as *komputerem* accounted for 20 (49%) spurious pre-1965 search results for *komputer*. 18 spurious results were caused by a library stamp, see figure 9[17] – strictly speaking, this is *not* an OCR error, such modern "inserts" are probably limited to few words ("library", "public", etc.).

In general, the older the text, the worse the quality of OCR. Some pre-1939 OCRed texts provided by digital li-

[15]"Kurjer Warszawski", 1834, no 226, p. 4, available at e-Biblioteka Uniwersytetu Warszawskiego, `http://ebuw.uw.edu.pl/Content/28059/directory.djvu`. The text quoted is a menu from one of the Warsaw restaurants.

[16]See Andrzej Bańkowski, *Etymologiczny słownik języka polskiego*, entry *KOMPUT*, p. 773. Interestingly, the dating given by this dictionary is 1846, 12 years after the fragment quoted.

[17]Silesian Digital Library, `http://www.sbc.org.pl/Content/13697/001.djvu`

Figure 8: A 1834 fragment with the word *komputem* misread

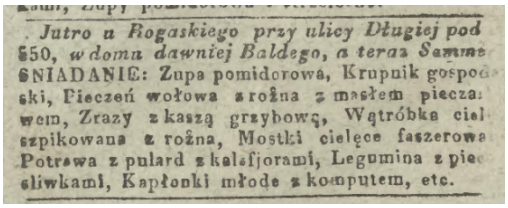


Figure 9: A library stamp on a 1848 book

braries are of very bad quality. Fragments like the following[18] are not unusual:

> Na łamach niektórych ogrodnkzych że 1&nie pomarańcze mag. 8t&nOWić dla cko trudno ulokować. W myjl sasady do pł9m roz,legły się głosy, daJłce wyrM o- nich konkurencjo, & konsument Z&^_ ut des wydaje siO korzyetnem, aby przy- bawom z powodu nadmiernego przywo- wne nie btdzie miał nic pneciw temu

5. Impact of noise on search results

In order to study the impact of noise on search results in a more systematic manner, we took a random sample of 318 lexemes from the PoliMorf morphological dictionary of Polish.[19] Only open-class words (nouns, verbs, adjectives and adverbs) were considered. A search engine built with the texts extracted from digital libraries was queried with each lexeme (the inflected forms were taken into account). The search results were ordered by date ascending, texts for which it was not possible to determine the publication/creation date were discarded. For each lexeme we checked manually which search result was the first one relevant and dated correctly (we will refer to this as P):

- $P = 1$ for 205 (65%) lexemes, i.e. for 205 lexemes the earliest result (as reported by the search engine) contained the occurrence of the lexeme in question,
- $P \in \{2, 3, 4, 5\}$ for 58 (18%) lexemes,
- $P > 5$ for 55 (17%) lexemes.

What were the main causes of $P > 1$?

1. It is very easy to get spurious results for *short words* in OCR, e.g.:
   - *slams* ($P = 77$!) was really *Stanisław*, *stanu*, *siamskiego*, etc. broken and/or misrecognised,
   - the earliest occurrences of *aria*, *arka*, *gar* and *tram* were mostly fragments of broken Latin texts (these were the Latin documents for which no language was specified in the metadata and which were assumed to be Polish by us),
   - search results for words such as *sarin* or *jazz* were caused by all sorts of errors (broken words, unrecognised Polish or foreign words, names).
2. Sometimes OCR errors resulted in spurious results for longer words as well, e.g. the first result for *przeciwrządowy* was really *przeciw rządowy*, *trawność – niestrawność*.
3. Simple homonymy (no OCR errors!) was involved in some cases, e.g. *profesorek* was returned for a fragment with *profesorka*.
4. The first result(s) were sometimes grossly misdated, e.g. the first 3 results for *wódka* were misdated (1391 instead of 1931, 1542 instead of 1942).

6. Conclusions

It is possible to extract a large, diachronic corpus of Polish from the documents stored in digital libraries. This corpus should be of interests not only to linguists or historians, but also to NLP practitioners, as it can be used for training and testing various NLP (automatic datation, OCR, informational retrieval) systems.

References

Garcia-Fernandez, A. et al. (2011): When Was It Written? Automatically Determining Publication Dates. Lecture Notes in Computer Science, 221-236.

Lyman, P. and Varian H.R. (2003): How Much Information?. SIMS, University of California at Berkeley, CA, US. published online: `http://www.sims.berkeley.edu/how-much-info-2003`

Michel, J.-B et al. (2011): Quantitative Analysis of Culture Using Millions of Digitized Books. Science 331, 176-182.

Podhajecka, M. (2010): Antedating headwords in the third edition of the OED: Findings and problems. In Anne Dykstra and Tannecke Schoonheim (eds.), *Proceedings of the XIV EURALEX International Congress (Leeuwarden, 6-10 July 2010). Leeuwarden: Fryske Akademy.* 1044-1064.

Wierzchoń, P. (2009): Fotodokumentacja 3.0. In P. Nowak, P. Nowakowski (eds.), *Language, Communication, Information.* 63–8.

Wierzchoń, P. (2010): Torując drogę teorii lingwochronologizacji [Leveling the ground for a theory of linguochronologization]. Investigationes Linguisticae, vol. XX.

[18] `http://kpbc.umk.pl/Content/74403/`
[19] `http://zil.ipipan.waw.pl/PoliMorf`