CMST 24(1) 7–19 (2018) DOI:10.12921/cmst.2018.0000004

Re-research.pl: where Humanities Meet Computer Science

Daniel Dzienisiewicz[1], **Łukasz Borchmann**[1], **Piotr Wierzchoń**[1], **Filip Graliński**[2]

[1] *Institute of Linguistics, Faculty of Modern Languages and Literatures*
Adam Mickiewicz University, Poznań, Poland
al. Niepodległości 4, 61-874 Poznań
E-mail: dzienis@amu.edu.pl, borch@amu.edu.pl, wierzch@amu.edu.pl

[2] *Department of Natural Language Processing, Faculty of Mathematics and Computer Science*
Adam Mickiewicz University, Poznań, Poland
Umultowska 87, 61-614 Poznań
E-mail: filipg@amu.edu.pl

Received: 20 March 2017; revised: 24 November 2017; accepted: 16 January 2018; published online: 31 March 2018

Abstract: The article discusses selected projects from the field of digital humanities realised by the Re-research.pl group. The group consists of researchers from the Institute of Linguistics and the Department of Natural Language Processing at Adam Mickiewicz University, Poznań, Poland. The projects discussed include *National Photocorpus of Polish*, *Discovermat*, *Korea, Koreans and 'Koreanity' in the digitised Polish press of the 20th century*, *Biography of the Nation*, *100,000 ministories*, *Gonito.net* and *50,000 words. Domain and chronologisation index*. However, the main focus of the article is the interdisciplinary popular-scientific blog Re-research.pl. The daily blog posts include texts on a variety of subjects, ranging from linguistics, history and folklore to computer science. Selected posts and categories of posts are discussed, such as chronologisational challenges, texts devoted to folklore and materials on the structure of text files. Apart from providing daily analyses, the blog promotes other projects and serves as a dialogue platform for representatives of various fields.

Key words: digital humanities, corpus linguistics, big data projects, linguochronologisation, photodocumentation, e-lexicography

I. INTRODUCTION

According to Klein and Gold, digital humanities (DH) have beyond doubt finally arrived as a field of study. The authors mention such initiatives within DH as digital archives, quantitative analyses, tool-building projects, visualisations of large image sets, 3D modelling of historical artefacts, "born digital" dissertations, hashtag activism, alternate reality games and mobile makerspaces [1]. However, as the authors suggest, it may be difficult at times to specify with sufficient precision what DH is actually concerned with, especially when referring to the broad understanding of the field, known as "big tent" DH. What is more, besides being a relatively new research area, since its inception DH has become a notion signifying various sorts of computational approaches in the humanities, which vastly increases its terminological vagueness. The goal we set for ourselves, however, is not an attempt to create yet another definition of DH or specify its scope. Our intention is to bring attention to certain endeavours undertaken by a group of scholars from Adam Mickiewicz University that can be considered a part of DH. Therefore, the aim of the present article is to describe selected projects realised by the Re-research.pl group, understood as DH-oriented and aspiring to contribute to the fast-

growing DH field. Their general aims will be outlined, and the tools and stages of work will be described. Furthermore, the content of the Re-research.pl website will be presented in a condensed fashion.

The label Re-research.pl refers to a group of scholars from Adam Mickiewicz University, Poznań, Poland. The members of the group are researchers from the Institute of Linguistics (IL – Faculty of Modern Languages and Literatures) and Department of Natural Language Processing (DNLP – Faculty of Mathematics and Computer Science) specialising in linguistics and computer science, respectively. The creators of Re-research.pl are Piotr Wierzchoń (IL), Filip Graliński (DNLP), Łukasz Borchmann (IL), Daniel Dzienisiewicz (IL), Rafał Jaworski (DNLP) and Szymon Kwapiszewski (IL). The first incentive for the group to cooperate on a regular basis was their joint work on *50,000 words. Domain and chronologisation index*, an ongoing project funded by the National Programme for the Development of Humanities and planned for the years 2015–2019. Work on that project resulted in close cooperation which included weekly meetings at IL and monthly seminars at DNLP.[1] As a result of mutual efforts, since 2015 the number of co-participating projects has increased significantly. Thus we shall first present selected projects currently being carried out by the authors.

II. SELECTED PROJECTS

As noted above, the ongoing collaboration between IL and DNLP has resulted in various DH-oriented projects, mostly originating in linguistics, but not restricted to that research area. The joint projects include, among others, *National Photocorpus of Polish*, *Discovermat*, *Korea, Koreans and 'Koreanity' in the digitised Polish press of the 20^{th} century*, *Biography of the Nation*, *100,000 ministories*, *Gonito.net* and *50,000 words. Domain and chronologisation index 1918–1939*. Descriptions of each of these initiatives are given below.

II. 1. National Photocorpus of Polish

The most popular form of lexicographic exemplification is plain-text transcript. Apart from the undoubted advantages of such a quotation method, the solution may be perceived as some kind of trade-off when considering readability, accessibility, simplicity, accuracy, and even the logistics of a documentation project.

Another way is to gather and present excerpts in the form in which they were originally published, namely as clippings from publications. This alternative (hereinafter referred to as photodocumentation) has been proposed and is constantly performed on a large scale by Wierzchoń [2, 3].

One of the distinguishing features of Re-research's projects is a photodocumentative approach to presenting quotations. This applies also to presenting lexicographic units, hence the name *Photocorpus*, where words are documented as a clipping from the original source, so that complete information about the form is preserved.

The National Photocorpus of Polish (NFJP)[2] was conceived as an extension of the *Depozytorium leksykalne języka polskiego* (Lexical Depository of Polish) – a 10-volume work authored successively by P. Wierzchoń, J. Wawrzyńczyk, A. Wawrzyńczyk, A. Zombirt, E. Małek and M. Iwanowski [4]. However, it also reflects the broader breakthrough in lexicography, which, from a discipline built around traditional, deeply philological instruments, has been transformed into an interdisciplinary field involving linguistics and computer science.

The main goal of NFJP was to describe around 250,000 lexical units, which would be enough to outperform all of the 20^{th}-century dictionaries of Polish. This goal was achieved and, moreover, the material gathered consists largely of words of which linguists had been unaware or which were perceived as later neologisms by leading derivative models of Polish. This is because of the rejection of an approach where entries are inherited from previous dictionaries, in favour of a corpus-driven method, beginning with the acquisition of printed books which are subsequently digitised and analysed.

The tools and means applied in subsequent stages also allow us to place the undertaking under the headings of electronic lexicography and digital humanities. This may be justified by, for example, the automation of phonematic transcription and morpheme segmentation, the development of which involved tests of multiple supervised, semi-supervised and unsupervised machine learning solutions. The final solutions based on Conditional Random Fields and Support Vector Machines classifier with linear kernel outperformed other methods, especially those based on unsupervised and semi-supervised machine learning techniques [5–8].

Another area that involves the electronic part of the NFJP project is the presentation layer and all of the features offered to the end-user, such as sophisticated search operators that can be used along with regular expressions to refine the results, and charts presenting word usage within desired periods (obtained through integration with the Discovermat system, cf. section 2.2).

In the course of the development of NFJP, other e-lexicographic projects have been derived from the original undertaking, namely the Great Photocorpus of 20^{th}-Century Vietnamese and the Great Photocorpus of Korean.

[1] A list of weekly DNLP seminars is available at https://psi.wmi.amu.edu.pl/seminar/ (accessed: March 9, 2017).
[2] http://nfjp.pl/ (accessed: March 18, 2017).

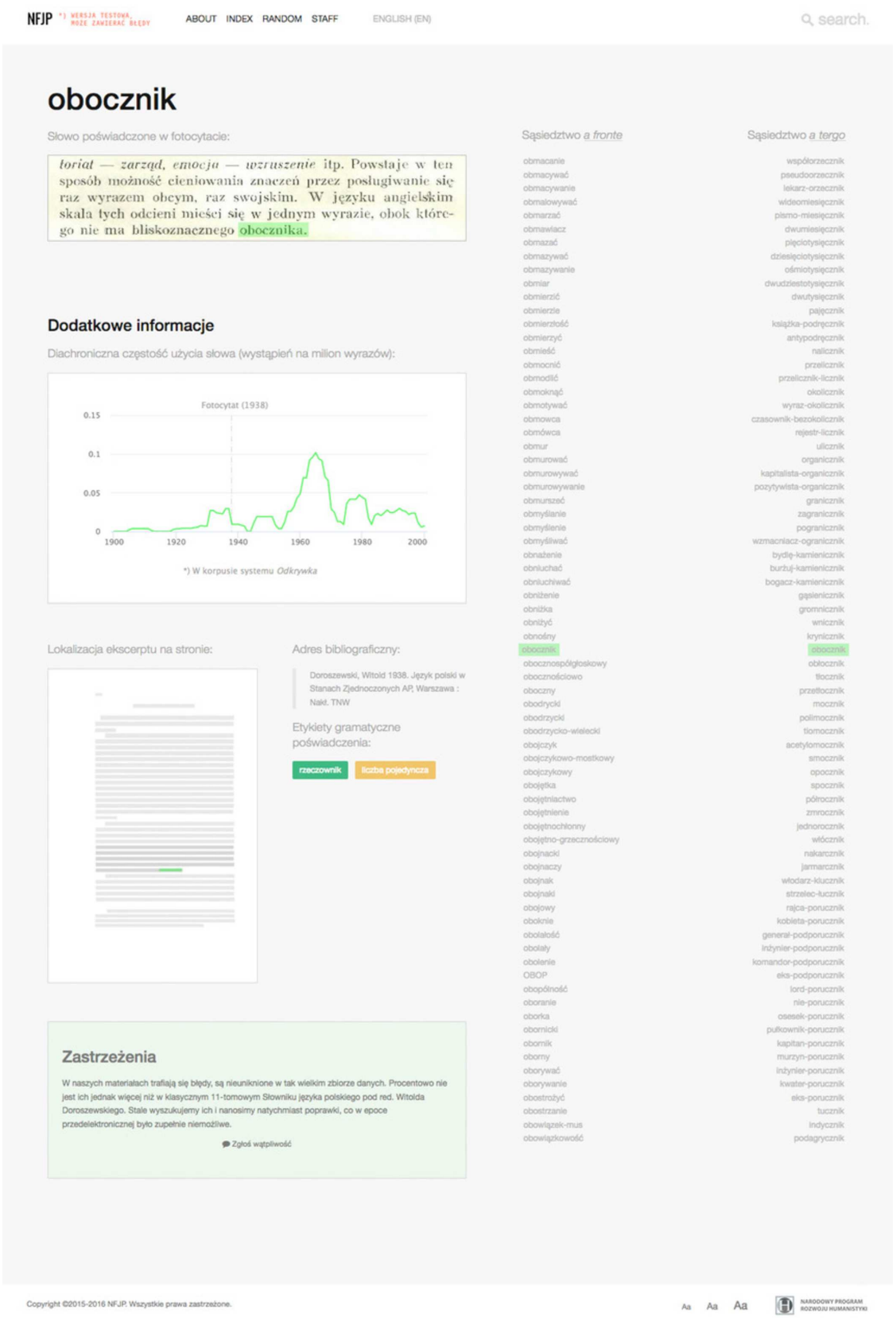


Fig. 1. Website of the National Photocorpus of Polish – view of the record *obocznik* 'collateral' (noun)

II. 2. Discovermat (Odkrywka)

'Discovermat' is an engine that allows one to perform full-text search in the largest diachronic corpus of Polish, consisting of hundreds of thousands of digitised texts from the 19^{th} and 20^{th} centuries(the collection also includes some texts from the 18^{th} and 21^{st} centuries).

It utilises the Apache Solr engine to index material consisting of 3.2M publications (19.7M pages, 18B words and 91B characters – as of March 20, 2017), gathered from Polish digital libraries and various other formal and informal initiatives, including digitisation performed at the Institute of Linguistics by the Re-Research group.

The primary aim of the analyses conducted with the use of the Discovermat corpus is to perform chronologisation and, when necessary, to antedate mischronologised units of language, and to present the data obtained in the form of scans taken from original printed matter. However, not only is Discovermat a useful tool for linguistic studies, but it also serves as a source of materials for analysing Polish history, culture and society with the use of historical texts. The system is assisted by automatically generated frequency graphs, making it possible to formulate research hypotheses in a fast and effective way [9–12]. Discovermat is equipped with a *Dossier* function which allows to export and upload scans on the Re-research blog in a fast manner. As far as the workflow is concerned, the materials are tagged as interesting or uninteresting from the perspective of the problem under investigation, and subsequently the photographic excerpts are cut out manually. Each stage of the process can be performed by a different person.

The system is currently available only to invited researchers; however, the first fully public version is being developed and will be launched soon.

II. 3. Korea, Koreans and 'Koreanity' in the digitised Polish press of the 20^{th} century

Propaganda – Topics unknown – Sensational Stories

This interdisciplinary big data project is directly linked with the Discovermat corpus, which serves as a source of textual material for the undertaking. It has as its goal the extraction and photodocumentation of over ten thousand pieces of Polish text pertaining to Korea. Not only will the references be collected, but they will also be analysed with regard to their appearance and placement in the texts from which they are extracted. Furthermore, the texts will be categorised thematically and geographically. Thematic categories are expected to include:

1. The Korean peninsula and "World History"
2. Polish–Korean relations in the second half of the 20^{th} century
3. Korean domestic affairs and internal politics
4. Korea as seen by foreigners
5. The economy, trade, technology
6. Culture and entertainment
7. Sport
8. Religion
9. Famous Koreans in the Polish press
10. The Korean natural environment.

The result of the work will be a large-scale database covering the period 1901–2000. Such a database will allow researchers to find original sociological and historical data on Korea, as well as trace the evolution of Polish people's views on Korea and its national identity in the 20^{th} century. This digital resource will include several tens of thousands of thematically grouped collections of publicly available and easily accessible entries. The entries will be made searchable thematically, chronologically, by frequency, etc.[3] Apart from the database, a series of studies on each of the above ten thematic fields will be conducted and published in the form of monographs, scientific articles and conference presentations.

II. 4. Biography of the Nation

This project aims to create a tool (*Automatyczny Biograf – Automated Biographer*) which will enable the user to find all references to people (i.e. names) in a massive collection of digitised Polish historical texts.[4] The idea is to greatly develop and add to the traditional concept of biographical dictionaries. Obviously no dictionary, no matter how thoroughly prepared, will ever record information on the masses of people who have worked for the sake of the nation's development and its interests through daily collective and individual efforts. The objective is to gather information not only about key figures (politicians, artists, entrepreneurs) but also about people who are not necessarily commonly known.

Automated Biographer will be created with the use of state-of-the-art computer techniques. The key technique in the project will be what is called entity recognition, which is a very challenging task, yet easier to perform than automated machine translation and speech recognition, for instance. Automated methods, however, will never be fully effective. Among the difficulties that arise in the course of work are cases of two people sharing the same name.[5] Other sources of problems are spelling reforms, variant forms of last names, Polish spellings of foreign names (e.g. *Schmidt/Szmit*), occurrences of last names with no first names, and difficulties in distinguishing proper nouns from common nouns (e.g. at the beginning of a sentence). Never-

[3] A demonstration version of the website, describing the goals of the project (in Polish and Korean) as well as presenting some of the excerpts already collected, is available at http://korea-xx.pl/ (accessed: March 18, 2017).

[4] For more information in Polish see http://www.biogramnarodu.pl/ (accessed: March 18, 2017).

[5] An interesting example is the case of Józef Stalin (Joseph Stalin), a shoemaker from Chabówka (Lesser Poland), whose name appeared in the Cracow press in 1926. See: http://re-research.pl/pl/post/2016-09-17-60019-stalin.html; http://re-research.pl/pl/post/2017-02-11-00046-jeszcze-o-jozefie-stalinie-szewcu-z-rabki.html (accessed: March 18, 2017).

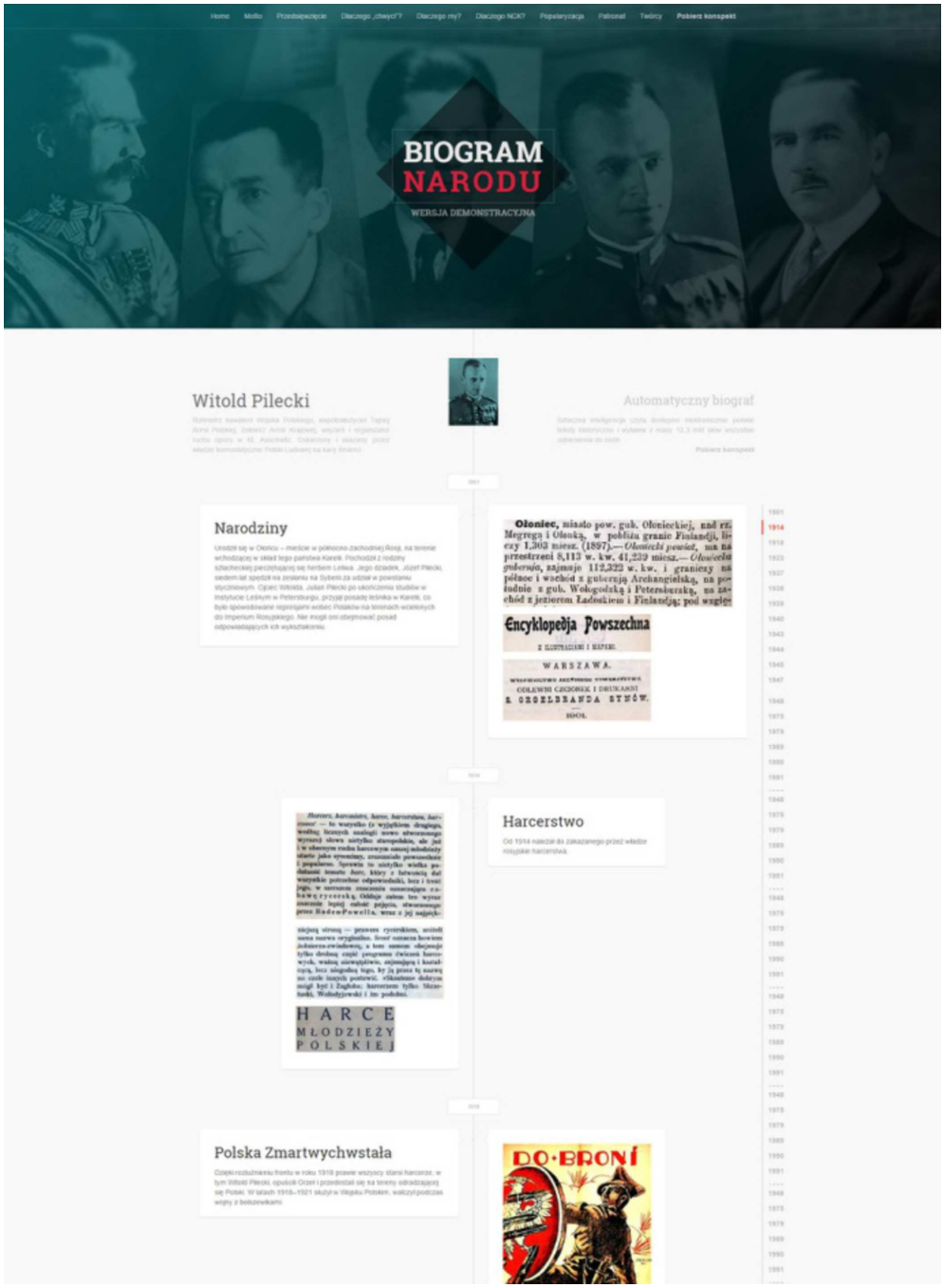


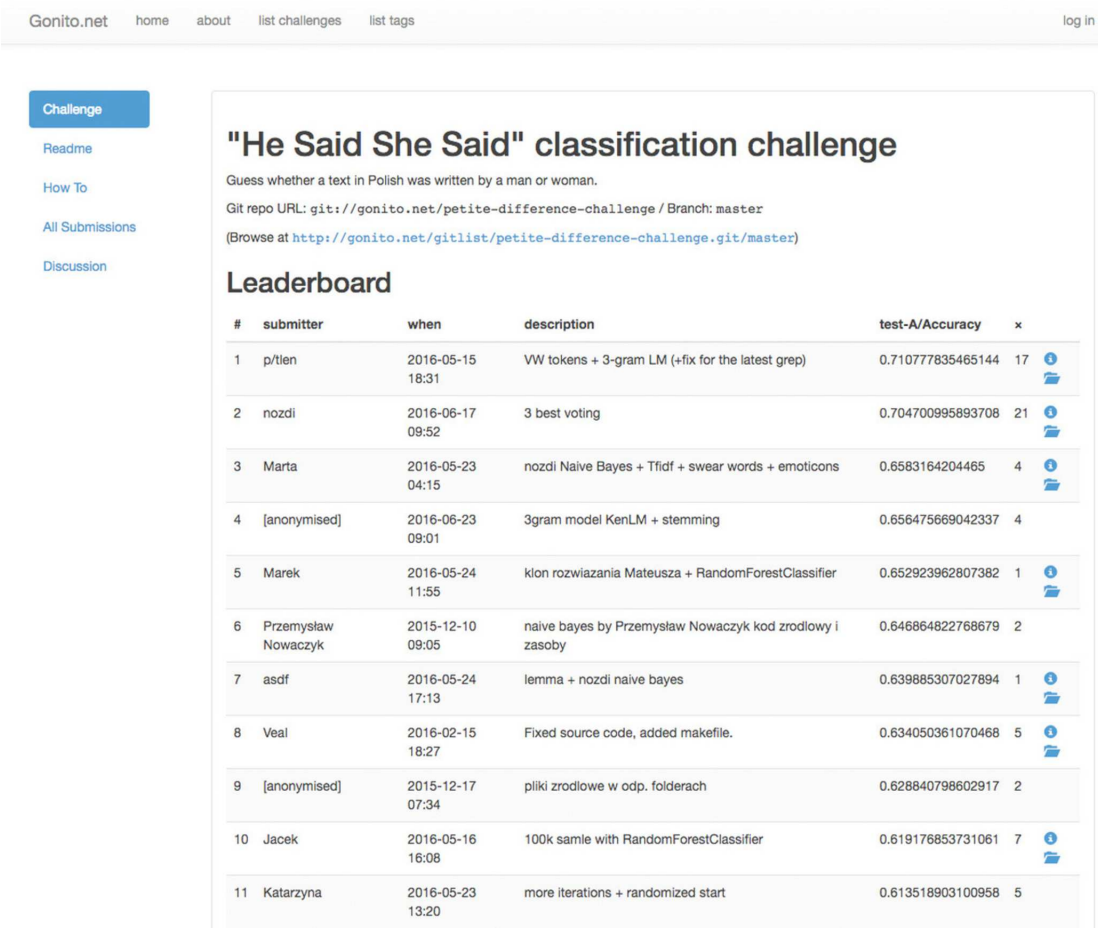
Fig. 2. Biography of the Nation – the view of Witold Pilecki's biography

theless, bearing in mind that the objective is to gather information on millions of people, only automated full-text search can be taken into consideration as an optimal method in this case. The results of the work will include:

1. automated searching for references to last names along with first names, titles, etc.;
2. an excerpt of text in the original photodocumentary form for each occurrence of a name;
3. automatically obtained information on time, space and social relations.

In sum, the aim is to present the history of any person mentioned in historical texts, starting from the earliest record to their obituary (if available).

Apart from the search engine, a dossier of people who struggled for Poland to regain its independence before 1918 is planned to be created. The materials collected with the help of the *Automated Biographer* will be manually verified,



#	submitter	when	description	test-A/Accuracy	×
1	p/tlen	2016-05-15 18:31	VW tokens + 3-gram LM (+fix for the latest grep)	0.710777835465144	17
2	nozdi	2016-06-17 09:52	3 best voting	0.704700995893708	21
3	Marta	2016-05-23 04:15	nozdi Naive Bayes + Tfidf + swear words + emoticons	0.6583164204465	4
4	[anonymised]	2016-06-23 09:01	3gram model KenLM + stemming	0.656475669042337	4
5	Marek	2016-05-24 11:55	klon rozwiazania Mateusza + RandomForestClassifier	0.652923962807382	1
6	Przemysław Nowaczyk	2015-12-10 09:05	naive bayes by Przemysław Nowaczyk kod zrodlowy i zasoby	0.646864822768679	2
7	asdf	2016-05-24 17:13	lemma + nozdi naive bayes	0.639885307027894	1
8	Veal	2016-02-15 18:27	Fixed source code, added makefile.	0.634050361070468	5
9	[anonymised]	2015-12-17 07:34	pliki zrodlowe w odp. folderach	0.628840798602917	2
10	Jacek	2016-05-16 16:08	100k samle with RandomForestClassifier	0.619176853731061	7
11	Katarzyna	2016-05-23 13:20	more iterations + randomized start	0.613518903100958	5

Fig. 3. Leaderboard of one of the challenges available on Gonito.net platform. Directory icon denotes that link to source code is available for open access

selected and described. The verification will help to reveal details unknown to historians, especially curiosities concerning a person's youth or facts which are not directly connected with their main activity. The pieces of information will be delivered in an attractive form, but with no distortion of historical accuracy. In addition, manual verification of search results will considerably improve the work of the *Automated Biographer* due to the application of machine learning techniques.

The *Biography of the Nation* cannot be a purely documentary project (performed by people and machines). The collected materials should inspire artists and scholars, such as poets, writers, historians and journalists. Thus a competition entitled *Łuk Tryumfalny ze Słów* (*A Triumphal Arc of Words*) will be run to select original works inspired by the *Biography of the Nation*. Suggested categories include poems, realistic novels, fantasy novels, screenplays, biographies, history books and popular-scientific books.

II. 5. 100,000 ministories

The aim of this project is to make a linguistic and sociocultural analysis of a corpus of Polish postcards sent in the years 1945–1989. The main idea of the initiative is to gather vast knowledge about Polish people's daily lives by building and examining the corpus of postcards. Scientific analysis will be performed on a collection of approximately 100,000 postcards, making it the largest digital resource of its kind in the world.[6] Every postcard will be scanned and transcribed. Subsequently, each component of the postcard text (greetings, wishes, signature, etc.) will be annotated ontologically to enable fast and effective quantitative measures of the structure of these texts. Moreover, an automated data

[6] There is no such collection for the Polish language. One can refer to hobbyists' collections such as http://www.polskie-pocztowki.com (accessed: March 18, 2017) and websites devoted to the topic: http://www.walkowiak.pl/poznan/pocztowki/index.html, http://przypadkipocztowkowe.blogspot.com/p/kolekcje.html (accessed: March 18, 2017). One can also find collections of postcards (www.pocztowki.plockie.pl/artykuly.html – accessed: March 18, 2017) with detailed descriptions of the photographs, but lacking analysis of the text. Postcards are also collected by Polish digital libraries; for example, the Digital Library of Wielkopolska contains over 400 postcards, although only the photographs are scanned (the obverse). None of these collections aims to present the texts, not to mention their linguistic and historical statistical analysis.

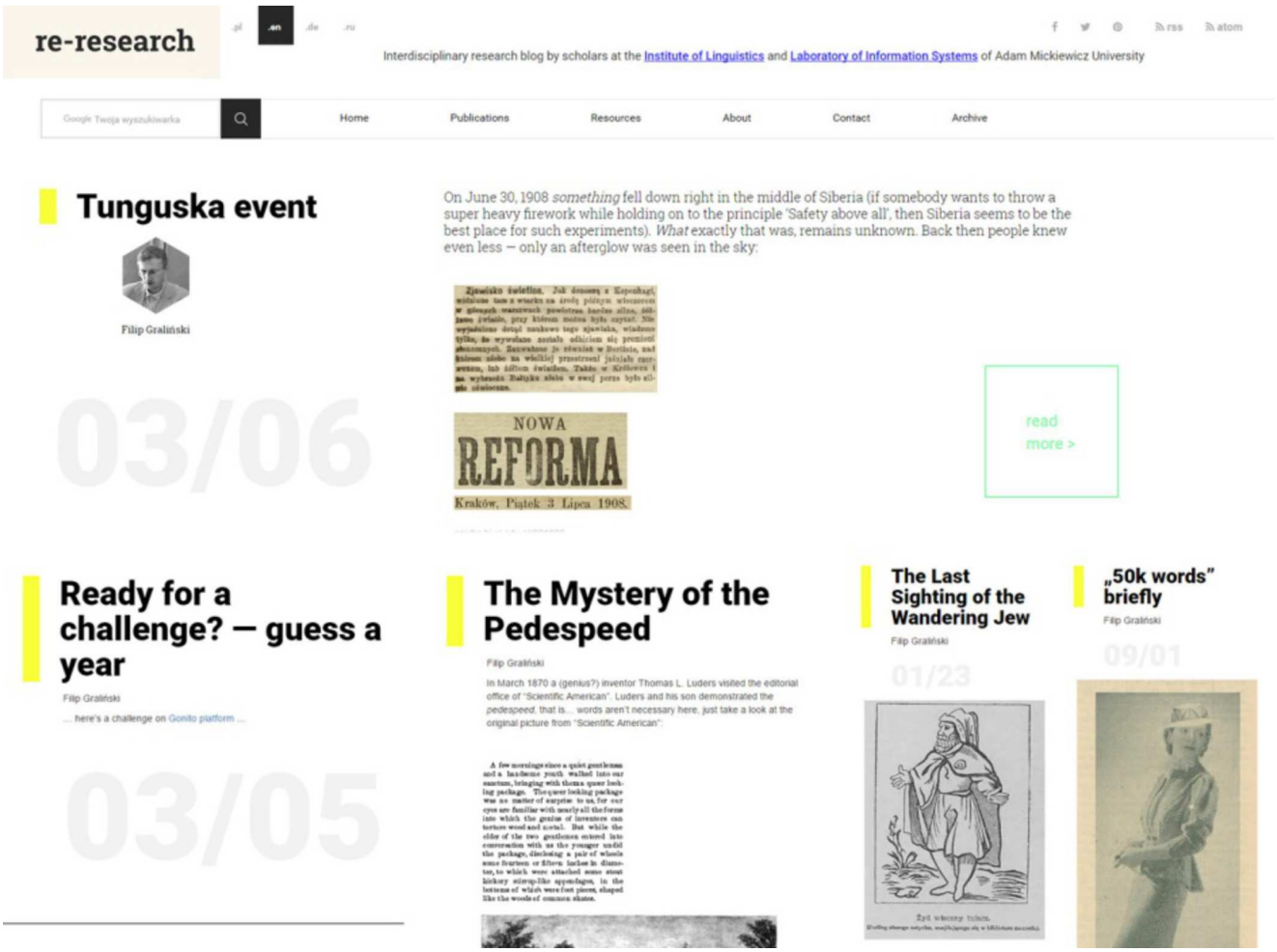


Fig. 4. Home page of the Re-Research.pl website

anonymisation tool (e.g. addressee and address data) will be created.

The records of cultural and socio-political changes which are expected to be found in the analysed material will be crucial both from a linguistic and a historical point of view. Postcard records are a true reflection of their times, as they include private comments. Knowledge will be gained about the everyday life of Polish people – the everyday reality of the 20^{th} century will be recreated (vacations abroad, increases in prices, food stamps, the coldest winter of the century, etc.). There will be three tangible results of the project:

An immediately obtainable result will be the largest, widely available, free of charge and, above all, searchable corpus of post-war Polish postcards. Such a corpus has not yet been created either in Poland or anywhere else in the world[7].

The methodological result will be a description of innovative operational solutions (designed as a result of close cooperation between information technology and linguistics) which led to the creation of the corpus. A monograph describing all of the operational steps will be published.

As for the long-term results, important for Polish culture and the humanities, a unique image of Polish society will be

[7] A demonstration version of the corpus is available at http://www.100000minihistorii.pl/ (accessed: March 18, 2017).

presented through the prism of an enormous collection of individual texts (postcard records). The result of this work will be a complex analysis of the accumulated material. As far as the linguistic component is concerned, holistic analyses of the corpus will be performed: the classification of texts by subject matter, analyses of the frequency of key words, phraseology and phrasematics, syntax, spelling and linguistic creativity (neologisms). Due to the fact that postcards are in most cases marked with a date (often to an accuracy of a single day), this information will make it possible to trace the evolution of all of the above aspects.

II. 6. Gonito.net

Gonito.net is an open Git-based [13, 14] platform for hosting machine learning challenges [15–17]. Its objective is to foster research competition, cooperation and reproducibility. Researchers are encouraged to compete in well-defined tasks by developing software tools, running them on provided test data and submitting the solutions to Gonito.net.

The non-toy challenges hosted so far at Gonito.net lie within the field of natural language processing (though Gonito.net could be used for any type of machine learning challenges). Most of the Gonito.net challenges are based on the data used in other projects related to Re-research.pl.

Although Gonito.net is intended for machine learning practitioners (i.e. computer scientists), the datasets available there are useful in themselves and might be of interest to a wider range of researchers.

A list of selected Gonito.net challenges is given below.

Sane words challenge. The task is to guess whether a given word is a correct Polish word in a given domain. 44,344 words were manually annotated for the training set of this challenge (11,061 for the test set). The words were taken from the corpus on which the Discovermat system is based. The selection criterion was to take words which can be found in domain-specific interwar publications and which are not listed in a large Polish lexicon.

RetroC temporal classification challenge. In this challenge, you are expected to create a system which is able to guess the publication year of a short Polish text. The dataset is a large (40,000 texts for the training set, 10,000 for the test set) Polish-language diachronic corpus, spanning two centuries (1814–2013). Again, the dataset was extracted from a larger corpus indexed within the Discovermat system. There also exists a similar but smaller challenge for Vietnamese.

Clipping death notices. The challenge is to create a system for identifying death notices in Polish newspapers. The training set consists of 3,191 publications (mostly dailies) and the test set of 532 publications. Methods from both natural language processing and computer vision should be used to tackle this challenge.

"He Said She Said" classification challenge. This is more a sociolinguistic than a diachronic challenge – the task is to guess whether a short Polish text was written by a man or a woman. The training set is very large for this classification task (3.6M texts).

Russian–Polish Open subtitles. This challenge requires a machine translation system for translating from Russian into Polish to be created. A parallel corpus of 3M sentences was provided for this task.

II. 7. 50,000 words. Domain and chronologisation index 1918–1939

50,000 words. Domain and chronologisation index 1918–1939 is a project having as its aim the creation of a thematic index of interwar Polish. 50,000 words and phrases will be retrieved from source texts from the period 1918–1939. The collected units will be categorised thematically and by sociocultural domain (e.g. religion, occultism, sports, philosophy, politics, mathematics, music, chemistry, everyday life) and presented in the photodocumentary form as clippings from texts. The list of entries and excerpts will be made publicly accessible as a digital resource.

The words are sought with the use of various methods, e.g. they are extracted from the available thematic dictionaries or with the use of word2vec models (a large vector model is trained on the large diachronic corpus of Polish texts and the words in the vicinity of the already known domain-specific words are found [18, 19]. The words are currently being accepted by linguists; however, the ultimate set of words will be accepted by domain experts.

Currently the list of entries is being prepared. The units obtained in the first part of the retrieval process are in most cases absent from the pre-World War II Polish lexicography. Several hundreds of thousands of texts from the system of Polish digital libraries, as well as texts collected separately due to their documentary value (textbooks, lists of subscribers, etc.), will be examined. An example of a similar work for English is the *Historical Thesaurus of the Oxford English Dictionary* [20]. The idea of *50,000 words...* is to create an analogous work for Polish, but comprising a shorter time period (1918–1939) and with entries attributed with a photograph of an original excerpt of text.

A series of ten monographs devoted to each domain and a publicly available electronic resource of thematically categorised entries will be published. The resource will be searchable alphabetically, thematically, chronologically, quantitatively, and grammatically (by part of speech, morphological structure, etc.). Every entry will be illustrated with a scan of an original piece of text where it appeared. The quotations will be selected in such a way as to best exemplify the typical textual environment of entries. Every monograph will include all entries attributed to a particular domain as well as excerpts for at least 50% of all entries from a particular field. The group of entries with the highest frequency will be subjected to additional analysis. The index will demonstrate the wealth of Polish interwar vocabulary pertaining to various areas of everyday life.

III. RE-RESEARCH.PL

Re-research.pl is an interdisciplinary popular-scientific blog and a promotional tool for the projects outlined above[8]. The main goal of the website is to popularise various types of studies based on historical texts, traditionally performed within the framework of such disciplines as corpus linguistics, linguochronologisation, folklore studies, history, sociology, natural language processing and machine learning. Furthermore, the site may potentially become a unifying environment and a dialogue platform for representatives of all of the aforementioned fields.

The blog comes in four language versions: Polish, English, German and Russian. All posts are originally written in Polish and only selected texts are translated into the other languages. The website presents a complete list of the authors' publications (see the *Publications* section) directly or indirectly related to their joint work, as well as a list of resources (the *Resources* section). The general idea of the blog and its authors' profiles are concisely described and displayed in the *About* section.

The blog was officially launched on September 1, 2016, and since then posts have been published daily. Over 200 texts have been published (202 texts as of March 20, 2017). The posts can be categorised into three main fields: linguistics, computer science and *varia*. Due to the large number of posts, only selected examples from each category will be discussed to illustrate the main ideas to which they relate.

III. 1. Linguistics

Posts on linguistics constitute the largest group and pertain to various aspects of the study of language, such as linguochronologisation, word etymology, semantics, orthographic fluctuations, etc. Certain texts come in series, which will be outlined here. Chronologically, the very first series on the blog was *Szynobus tygodnia* (*Railbus of the week*), launched in September 2016. Its aim is to present attestations of occurrences of words taken from the list of entries of *Obserwatorium Językowe Uniwersytetu Warszawskiego*[9] (OJ UW, *Language Observatory of the University of Warsaw*) older than the year 2000 in the form of excerpts from texts. The goal of OJ UW is to record new words which entered the Polish language after the year 2000. However, the project has been criticised due to numerous chronological mistakes discovered in the entries [21, 22]. For instance, the post *Szynobus tygodnia 7* displays a record of usage of the word *przedkonferencja* (*preliminary conference*) from 1876 in its contemporary meaning[10]:

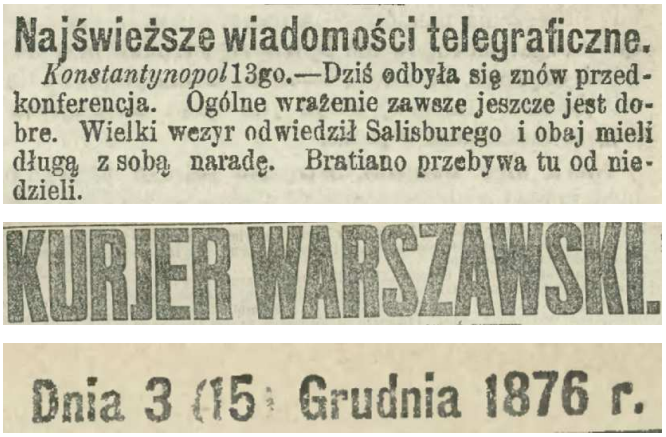

Najświeższe wiadomości telegraficzne.
Konstantynopol 13go.—Dziś odbyła się znów przedkonferencja. Ogólne wrażenie zawsze jeszcze jest dobre. Wielki wezyr odwiedził Salisburego i obaj mieli długą z sobą naradę. Bratiano przebywa tu od niedzieli.
KURIER WARSZAWSKI.
Dnia 3 (15) Grudnia 1876 r.

Fig. 5. The earliest record of the word *przedkonferencja*

Another series entitled *Bez komentarza* (*No comments*) aims to present frequency graphs for words and phrases with no additional explanations, in order for readers to infer conclusions for themselves. The content of posts often reflects extralinguistic facts, such as cultural and sociopolitical changes. For example, *Bez komentarza 14*[11] and *Bez komentarza 16*[12] display graphs for queries on *walentynki* (*Saint Valentine's Day*), *święty Walenty* (*Saint Valentine*), *wzornictwo*, *design* and *dizajn* (*design*), indicating the prevalence of foreign concepts (also expressed through the choice of linguistic means) over native ones in Polish texts at the turn of the 20^{th} and 21^{st} centuries:

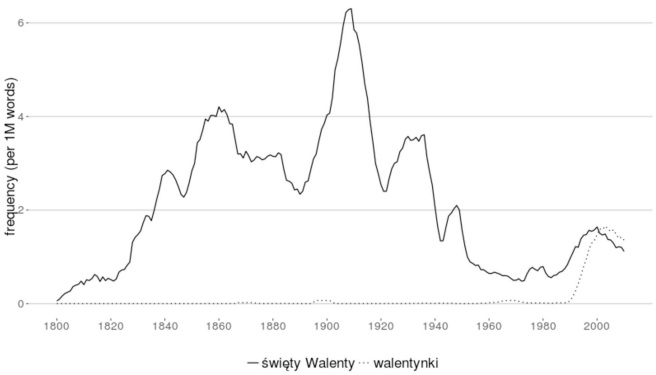


Fig. 6. Frequency graph for queries on *święty Walenty* and *walentynki*

[8] http://re-research.pl/pl/ (accessed: March 18, 2017).
[9] http://nowewyrazy.uw.edu.pl/ (accessed: March 18, 2017).
[10] http://re-research.pl/pl/post/2016-10-24-60044-szynobus-tygodnia-7.html (accessed: March 18, 2017). In total, around 40 occurrences of the word przedkonferencja from before 2000 were found.
[11] http://re-research.pl/pl/post/2017-02-14-00047-bez-komentarza-14.html (accessed: March 18, 2017).
[12] http://re-research.pl/pl/post/2017-03-03-00049-bez-komentarza-16.html (accessed: March 18, 2017).

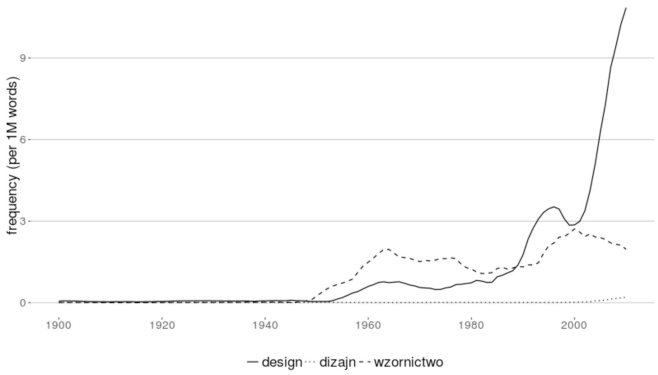


Fig. 7. Frequency graph for queries on *design*, *dizajn* and *wzornictwo*

Fig. 9. Frequency graph for queries on *bizantyjski* and *bizantyński*

Na krzyż (*Crosswise*), in turn, is a series displaying quantitative information regarding orthographic variants of words in the form of frequency graphs in a period in which one form is going out of use and the other is gaining popularity. Such graphs form a characteristic "cross" pattern, as with the graphs for the variant forms *nadto* and *ponadto* (*moreover*)[13]:

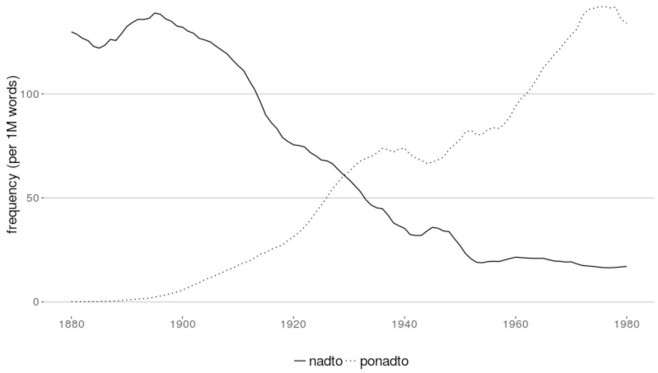


Fig. 8. Frequency graph for queries on *nadto* and *ponadto*

As concerns the frequency of orthographic variants, analyses are also performed over a longer period of time (e.g. 200 years). The following example shows the frequency of the forms *bizantyjski/bizantyński* (*Byzantine*) in the years 1800–2010[14]:

A series *Koreanizmy w polszczyźnie* is devoted to searching for words of Korean origin in Polish historical texts. So far the following units have been described and photodocumented: *Suwon*, *Samsung*, *taekwondo*, *kimczi* and *dżucze*.[15]

The blog presents other types of linguistic content as well, e.g. etymological hypotheses formulated on the basis of textual evidence. Certain 19^{th}-century texts may indicate that the Polish idiom *udawać Greka* (*to play possum*, literally: to pretend to be a Greek) might have been inspired by the biography of Titus Albucius, a Roman hellenophile, who nearly completely adapted to Greek culture after he was sentenced for crimes that he had committed in Sardinia.[16]

The authors also organise chronologisational challenges under the motto *Kto da wcześniej* (*Who'll go earlier*). So far two challenges, regarding the idiom *na pół gwizdka*[17] and the word *komputer*, have been announced[18]. The rules for challenges are to antedate a given word or phrase by finding an earlier attestation of their appearance in texts. The winner receives a prize of 5 zlotys for every occurrence older by a year than the previously formulated hypothesis.

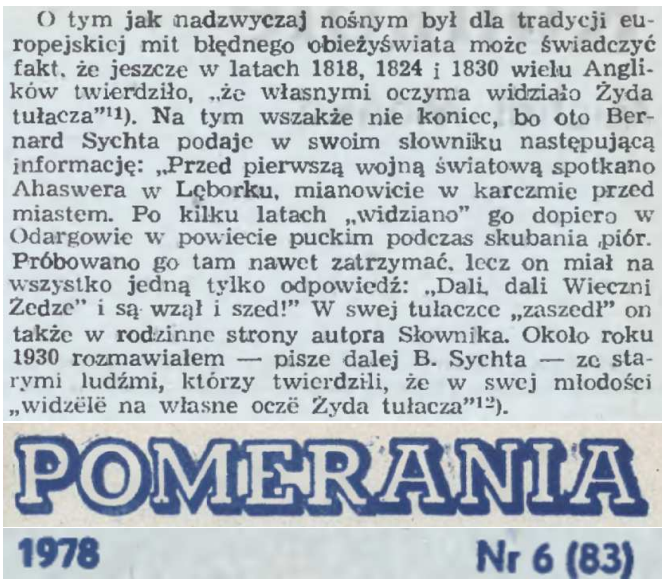
III. 2. 3.2. Computer science

The IT side of the website includes announcements about the challenges held on the Gonito.net platform (*He said, she said*, *Sane words*[19]) as well as the series *Anatomia pliku tekstowego* (*Anatomy of a text file*) describing the structure of text files, which serves as a source of basic information for laypeople. Moreover, a manual on how to keep abreast with the user's favourite Internet websites has been published[20].

[13] http://re-research.pl/pl/post/2017-01-20-60098-na-krzyz.html (accessed: March 18, 2017).
[14] http://re-research.pl/pl/post/2016-10-05-60029-bizantyjski.html (accessed: March 18, 2017).
[15] http://re-research.pl/pl/post/2017-03-06-00051-koreanizmy-w-polszczyznie–suwon.html; http://re-research.pl/pl/post/2017-03-07-60128-skoro-o-korei-mowa.html; http://re-research.pl/pl/post/2017-03-11-60132-taekwondo.html; http://re-research.pl/pl/post/2017-03-13-00049-koreanizmy-w-polszczyznie–kimczi.html; http://re-research.pl/pl/post/2017-03-14-00052-koreanizmy-w-polszczyznie–dzucze.html (accessed: March 18, 2017).
[16] http://re-research.pl/pl/post/2016-10-21-60042-tytus-albucjusz-udawal-greka.html (accessed: March 18, 2017).
[17] http://re-research.pl/pl/post/2016-09-13-00009-na-pół-gwizdka.html (accessed: March 18, 2017).
[18] http://re-research.pl/pl/post/2016-09-21-60015-komputer.html (accessed: March 18, 2017).
[19] http://re-research.pl/pl/post/2016-11-16-00021-seks-wyzwanie.html; http://re-research.pl/pl/post/2016-12-11-80003-zdrowe-slowa.html; http://re-research.pl/pl/post/2017-01-26-00043-na-froncie-plci.html (accessed: March 18, 2017).
[20] http://re-research.pl/pl/post/2016-10-04-00011-jak-sledzic-re-research.html (accessed: March 18, 2017).

III. 3. 3.3. Folklore

The posts on folklore are largely aimed at presenting records of (urban) legends discovered in historical texts. An example is the post on Ahasverus (the Wandering Jew). According to a mediaeval legend, Ahasverus was the man who rushed Christ on his way to Calvary and was subsequently banished to Eternal Wandering. By all accounts, the last sighting of the Wandering Jew was in 1868 in Salt Lake City. However, a later occurrence of a story about Ahasverus has been found in a Polish text from 1978[21]. The story, which allegedly took place in Kashubia, is presented below:



O tym jak nadzwyczaj nośnym był dla tradycji europejskiej mit błędnego obieżyświata może świadczyć fakt, że jeszcze w latach 1818, 1824 i 1830 wielu Anglików twierdziło, „że własnymi oczyma widziało Żyda tułacza"[11]). Na tym wszakże nie koniec, bo oto Bernard Sychta podaje w swoim słowniku następującą informację: „Przed pierwszą wojną światową spotkano Ahaswera w Lęborku, mianowicie w karczmie przed miastem. Po kilku latach „widziano" go dopiero w Odargowie w powiecie puckim podczas skubania piór. Próbowano go tam nawet zatrzymać, lecz on miał na wszystko jedną tylko odpowiedź: „Dali, dali Wieczni Żedze" i są wziął i szed!" W swej tułaczce „zaszedł" on także w rodzinne strony autora Słownika. Około roku 1930 rozmawiałem — pisze dalej B. Sychta — ze starymi ludźmi, którzy twierdzili, że w swej młodości „widzëlë na własne oczë Żyda tułacza"[12]).

POMERANIA

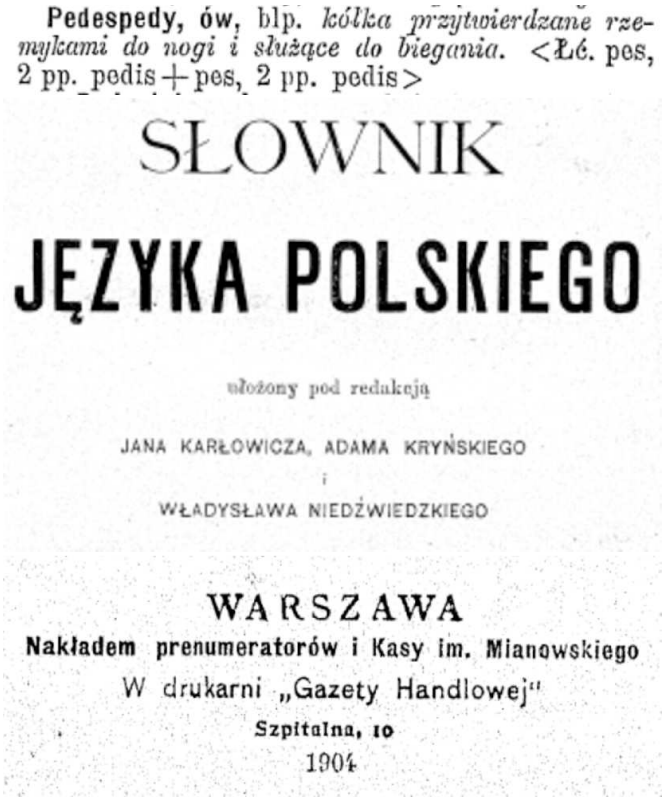
1978 Nr 6 (83)

(*The proof of the popularity of the story of the eternal wanderer in European tradition is the fact that in 1818, 1824 and 1830 many Englishmen claimed that "they had seen the Wandering Jew with their very own eyes". However, it is not over yet, as Bernard Sychta provides the following piece of information in his dictionary [of Kashubian dialects]: "Before World War I Ahasverus was seen in Lębork, namely in an inn right next to the town. After a few years he was seen picking at feathers in Odargowo, Puck county. Some people even tried to stop him, but he had only one answer to all their questions: 'Come on, the Eternal Jew' and then he was gone." In his eternal wandering he also visited the homelands of the author of the dictionary. Around 1930 I talked, writes B. Sychta, with old people who claimed that "they had seen the Wandering Jew with their own eyes in their youth."*)

Fig. 10. The record of the legend of the Wandering Jew

III. 4. Varia

Texts classified as *varia* cover a broad thematic spectrum; for example, notes on the history of peculiar inventions, such as the pedespeed (quasi roller skates invented by Thomas L. Luders in 1870), which was not commonly known at the time (let alone popular). The name, however, was included in *The Warsaw Dictionary* as a headword[22]:



Pedespedy, ów, blp. *kółka przytwierdzane rzemykami do nogi i służące do biegania.* <Łć. pes, 2 pp. pedis + pes, 2 pp. pedis>

SŁOWNIK

JĘZYKA POLSKIEGO

ułożony pod redakcją

JANA KARŁOWICZA, ADAMA KRYŃSKIEGO

i

WŁADYSŁAWA NIEDŹWIEDZKIEGO

WARSZAWA

Nakładem prenumeratorów i Kasy im. Mianowskiego

W drukarni „Gazety Handlowej"

Szpitalna, 10

1904

Fig. 11. The record of the word *pedespedy*

Other examples of humorous content are records of jokes from the interwar period[23]:

— Marszałek Piłsudski smaruje wąsy *dyktatuarem*..

— Centrolew twierdzi, że sanacja mu już *beckiem* wyłazi... —

Nr. 41. 30 gr.

TYGODNIK LITERACKI, HUMORYSTYCZNO-SATYRYCZNY

UŚMIECH

Październik Rok IV

W Ameryce 10 centów.

Fig. 12. The record of interwar jokes

[21]http://re-research.pl/en/post/2017-01-23-00040-the-last-encounter-with-the-wandering-jew.html (accessed: March 18, 2017).
[22]http://re-research.pl/en/post/2017-01-24-00034-the-mystery-of-pedespeed.html (accessed: March 18, 2017).
[23]http://re-research.pl/pl/post/2016-11-13-60057-miedzywojenne-zarty.html (accessed: March 18, 2017).

Moreover, the blog is used as a platform where conference reports are published. The authors have published four reports from the following conferences: *19^{th} International Conference on Text, Speech and Dialogue*, the 7^{th} edition of *Kultury Wschodniosłowiańskie – oblicza i dialog (Eastern Slavic Cultures – Faces and Dialogue), the* 3^{rd} edition of the *DARIAH-PL* conference, and the *International Conference on Asian Linguistics.*

IV. CONCLUSION

The initiatives described herein are aimed at achieving results of massive proportions. The combination of empirical and computational research makes it possible to provide real documented information on a wide range of topics, such as Polish language, history, culture, society, etc. in a fast and effective way. The final versions of the resources will hopefully contribute to increasing knowledge on various aspects of Polish history over two centuries. Not only will descriptions of the analysed materials be provided, but also tools and collections (indexes, corpora, etc.) which will serve as a basis for conducting further studies on a variety of subjects. The open access principle will allow any researcher and non-researcher to acquire information and perform various analyses individually.

In the context of the other initiatives, the Re-research.pl website serves as a promotional tool for the scientific activity of the group, as well as serving as a dialogue platform both for its authors and for other researchers representing a variety of fields.

References

[1] L.F. Klein, M.K. Gold, Digital Humanities: *The Expanded Field*, [in:] M.K. Gold, L.K. Klein, *Debates in the Digital Humanities*, University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/2.

[2] P. Wierzchoń, *Fotodokumentacja, chronologizacja, emendacja: teoria i praktyka weryfikacji materiału leksykalnego w badaniach lingwistycznych*, Instytut Językoznawstwa Uniwersytetu im. Adama Mickiewicza, Poznań 2008.

[3] P. Wierzchoń, *Dlaczego fotodokumentacja? dlaczego chronologizacja? dlaczego emendacja?: instalacja gazowa, parking podziemny i „odległość niezerowa"*, Instytut Językoznawstwa Uniwersytetu im. Adama Mickiewicza, Poznań 2009.

[4] P. Wierzchoń, *Depozytorium leksykalne języka polskiego. Nowe fotomateriały z lat 1901–2010*. Tom I., Uniwersytet Warszawski, Instytut Lingwistyki Stosowanej – BEL Studio, Warszawa 2010.

[5] T. Ruokolainen, O. Kohonen, K. Sirts, S. Grönroos, M. Kurimo, S. Virpioja, *A comparative study of minimally supervised morphological segmentation*, Computational Linguistics **42**(1), 91–120 (2016).

[6] J. Goldsmith, *Unsupervised learning of the morphology of a natural language*, Computational Linguistics **27**(2), 153–198 (2001).

[7] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 1 edition, 2000.

[8] M. Creutz, K. Lagus, *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0.*, Technical Report A81, Helsinki 2005.

[9] F.F. Graliński, *Polish digital libraries as a text corpus*, [in:] Z. Vetulani, H. Uszkoreit (eds.) *Proceedings of 6th Language & Technology Conference,* Fundacja Uniwersytetu im. Adama Mickiewicza, p. 509-513, 2013.

[10] F. Graliński, Folklorystyka 2.0, in: P. Grochowski (ed.) *NETLOR. Wiedza cyfrowych tubylców,* Wydawnictwo Naukowe Uniwersystetu Mikołaja Kopernika, p. 119-132, 2013.

[11] D. Dzienisiewicz, P. Wierzchoń, *On the Japaneseness of Polish: a Linguochronological Approach,* [in:] M. Iwanowski (ed.) *Opuscula Iaponica et Slavica 3*, BEL Studio, p. 53-76, 2016.

[12] P. Wierzchoń, *Słownictwo lat 30. XX wieku w obrazach i liczbach*, BEL Studio, Warszawa 2016.

[13] K. Ram, *Git can facilitate greater reproducibility and increased transparency in science*, Source Code for Biology and Medicine **8**(1), 1–8 (2013).

[14] D. Spinellis, *Version control systems*, Software, IEEE **22**(5), 108–109 (2005).

[15] R. Jaworski, Ł. Borchmann, P. Wierzchoń, *Gonito.net – Open Platform for Research Competition, Cooperation and Reproducibility*, [in:] B. António, N. Calzolari, K. Choukri (eds.) *Proceedings of the 4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*, pp. 13-20, 2016. http://4real.di.fc.ul.pt/wp-content/uploads/2016/04/4REALWorkshopProceedings.pdf.

[16] F. Graliński, Ł. Borchmann, P. Wierzchoń, *'He Said She Said' – a Male/Female Corpus of Polish*, [in:] N. Calzolari, K. Choukri, T. Declerck, et al. (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),* European Language Resources Association (ELRA), 2016. http://www.lrec-conf.org/proceedings/lrec2016/pdf/905_Paper.pdf.

[17] F. Graliński, R. Jaworski, Ł. Borchmann, P. Wierzchoń, *Vive la Petite Différence! Exploiting Small Differences for Gender Attribution of Short Texts*, [in:] *Lecture Notes in Artificial Intelligence,* pp. 54-61, 2016.

[18] T. Mikolov, K. Chen, G. Corrado, J. Dean, *Efficient Estimation of Word Representations in Vector Space*, 2013. https://arxiv.org/pdf/1301.3781.pdf.

[19] Ł. Borchmann, F. Graliński, R. Jaworski, P. Wierzchoń, *A semi-automatic method for thematic classification of documents in a large text corpus*, [in:] F. Mambrini, M. Passarotti, C. Sporleder (eds.) *Proceedings of the Workshop on Corpus-Based Research in the Humanities* (CRH), 2015.

[20] C. Kay, J. Roberts, M. Samuels, I. Wotherspoon, *Historical Thesaurus of the Oxford English Dictionary*, Oxford University Press, Glasgow 2009.

[21] J. Wawrzyńczyk, *300 tysięcy czy milion(y)? O stanie zasobów słownictwa polskiego w dniu 31 grudnia 2000 r.*, Mila Hoshi, Warszawa 2015.

[22] P. Wierzchoń, F. Graliński, *Z kart historii „parcia na" neologizmy*, Poradnik Językowy **4**, 110–129 (2016).

Daniel Dzienisiewicz is a PhD student and lecturer at the Institute of Linguistics, Adam Mickiewicz University in Poznań, Poland. His main research interests include phrasematics and phraseology, corpus linguistics, lexicology, word-formation, idiolect studies and English phonetics.

Łukasz Borchmann is a software developer and PhD student at the Institute of Linguistics, Adam Mickiewicz University. His research interests include various fields within and outside the humanities, especially machine learning and natural language processing.

Piotr Wierzchoń is the Director of the Institute of Linguistics, Adam Mickiewicz University in Poznań, Poland. Among others, his works are devoted to corpus linguistics, phrasematics, the chronologisation of the Polish lexical inventory of the 19th, 20th and 21st centuries, lexicography and the grammar of action. He is the author of the theory of linguochronologisation (TLCH).

Filip Graliński is an assistant professor at the Department of Natural Language Processing, Adam Mickiewicz University in Poznań, Poland. His main research interest is computational linguistics, that is, creating and processing diachronic corpora, natural language processing, syntactic analysis, lexicography, machine translation, named entity recognition and extracting linguistic data from the Internet. Furthermore, he is interested in contemporary folklore (mainly urban legends). He is the author of the book Znikająca nerka. Mały leksykon współczesnych legend miejskich (The Disappearing Kidney. A Brief Lexicon of Contemporary Urban Legends).

DOI:10.12921/cmst.2018.0000004