ВОПРОСЫ ТЕРМИНОЛОГИИ

Я. Вавжинчик, доктор филологических наук, заслуженный профессор факультета неофилологии кафедры прикладной лингвистики Варшавского университета, Польша; e-mail: j.wawrzynczyk@uw.edu.pl

Петр Вежхонь, доктор филологических наук, профессор, директор Института лингвистики Университета имени Адама Мицкевича в г. Познани, Польша; e-mail: wierzch@amu.edu.pl

КАК РЕВОЛЮЦИОНИЗИРОВАТЬ ТЕХНОЛОГИЮ СОСТАВЛЕНИЯ СЛОВАРЕЙ[1]

В статье описываются методы создания одноязычных словарей на примере формирования Национального фотокорпуса польского языка.

Ключевые слова: лексикография, одноязычный словарь, корпусная лингвистика, польский язык.

Jan Wawrzyńczyk, Dr. Sc. (Philology), Professor Emeritus, Faculty of Modern Languages, Department of Formal Linguistics, Warsaw University, Poland; e-mail: j.wawrzynczyk@uw.edu.pl

Piotr Wierzchoń, Dr. Sc. (Philology), Professor, Director of the Institute of Linguistics (IJ), Adam Mickiewicz University in Poznań (UAM), Poland; e-mail: wierzch@amu.edu.pl

HOW TO REVOLUTIONIZE THE TECHNOLOGY OF DICTIONARY-MAKING[1]

The article describes various methods of making monolingual dictionaries and is a case study of the National Photocorpus of the Polish Language.

Key words: lexicography, monolingual dictionary, corpus linguistics, Polish language.

1. Introduction

Modern dictionaries tend to increasingly include information about the time of coinage of entries. In the present article we would like to describe one effective procedure leading to the creation of a monolingual documentary dictionary that presents the dates of occurrence of its entries. As an example, the procedure of creating the *National Photocorpus of the Polish Language* (www.nfjp.pl) will be presented.

[1] Grant nr 0014/NPRH3/H11/82/2014: National Photocorpus of the Polish Language.

The aim of the project was to create the world's largest lexical collection of words and word combinations for the Polish language of the 20th century, with attestation by citations.

1) The key distinguishing principle will be that every excerpt is documented photographically, that is, in the same form in which it appeared in print.

2) Lexical observation will cover the period 1901–2000, that is, it will include texts published during that time.

3) Every excerpt will be precisely localized, that is, a record will be given of the title of the document from which it is taken.

4) Every excerpt will be given a precise chronological characterization, that is, a record will be given of the date on which it appeared in the text.

5) A new feature of the project compared with previous observations of 20th-century vocabulary will be that the work will result in the world's largest collection of units whose existence was not previously known, that is, units which have not previously been searched for and recorded.

This paper aims to show how important it is for chronologization, particularly when concerned with the vocabulary of the twentieth century, to obtain a broad degree of contact with collected materials, and how important it is to catalogue textual attestations for as many words as possible.

In this article we shall consider the isolation of linguistic units for the purpose of linguochronologization, and the performance of their chronologization based on a mass of texts actually produced and recorded in the twentieth century.

This is a short presentation of methods of searching for language units and determining their chronologization. It is also a presentation about the need to produce documentation, particularly photodocumentation – that is, to present specimens of language units in the form of photographic documentation on a large scale. The presentation also covers excerptology carried out for the purposes of linguochronologization in the 21st century, and also specifically excerptology carried out using photodocumentation.

2. Chronologization

Linguistic chronologization, also known as **linguochronologization**, involves assigning chronological information to linguistic objects (words, syntagms, etc.). The information assigned is simply a value of the parameter TIME, such as a year, a month, or even a specific minute. The level of chronological precision of the information results from the time parameter presented in a given description, that is, in a given theory. In

the literature on the subject, the parameter usually takes yearly values (1901, 1939, 1953, etc.).

The result of work on linguochronologization is called linguochronography. In order to have something to chronologize, that something must first be found and isolated. The isolation of units from texts is the task of excerptology. Hence linguochronologization lays down the theoretical framework for an undertaking oriented towards excerption and chronologizing, that is, it establishes the linguistic object being subject to chronologization, the localization parameters by which it is characterized, the excerption directives used to isolate that derivate from actually available texts, and so on. Linguochronography involves the creation of dictionaries serving to record dating pairs, that is, units of lexicographical description consisting of a head object, such as a word, together with its assigned date.

A particular role in these investigations is played by the procedure for determining the boundary dates for particular linguistic units, understood as the chronological moment at which a unit of the language appeared in the system, initially as what is known as a neologism, or sometimes *neonym*. These two terms have the same meaning, but *neonym* forms a series with such terms as *antonym*, *synonym*, *hyponym* and so on. The basis for work on linguochronologization and linguochronography is real, existing texts.

Why chronologization? Because by assigning years to linguistic objects, we are able to state which linguistic objects do not occur before a boundary date assumed in a given description or model. This in turn helps us to describe the development of the systems of a language, particularly the lexical and morphological systems. What we mean here is a description of the dynamics of changes and so on. We may be interested, for example, in Greek and Latin morphemes like *super-, hyper-, anti-, bio-, geo-,* and how productive each of them was in particular periods. On the other hand, chronologization of linguistic objects also lets us identify phenomena that cease to appear after a given time. For example, one may ask when such units of language as *Rain Napper*, *Belly Timber*, etc. went out of use. Then another question is why it was that those units stopped being used.

A particularly important problem in chronologization is the correct dating of vocabulary from the 20th century. It is particularly significant to find out which words date from before the war, and which appeared only after it ended. In Russia a similar boundary is marked by the October Revolution of 1917, while in Korea it is the end of the Korean War in 1953, and in Vietnam the year 1975. As we can observe, political watersheds, wars and revolutions, including technical revolutions, very often have a huge impact on a language's vocabulary.

The **linguistic basis** for identifying a new unit of language is **textual attestation**, and not the absence of lexicographic attestation in a general

dictionary of the language. Dictionaries are simply not able to record the day-to-day development of vocabulary, and indeed they generally feature huge gaps and omissions. Lexicographers fail to record a larger number of words than they record. For this reason, in work on chronologization, it is texts, not dictionaries, that are most important. The fact that a word does not appear in a dictionary does not mean that the word did not occur earlier than that date.

For example, the need to take account of texts is acknowledged by the authors of the *Historical Thesaurus of the Oxford English Dictionary* [Kay et al., 2009]. The dates given in that work are based on source materials (i.e. natural texts), because only chronological information obtained from original texts is relevant to the dating of particular words. As examples, we have the dating of the word *crankery* to 1884, *crankiness* to 1870, *crankism* and *crankness* to 1890, *eccentricity* to 1657, *faddishness* to 1884, *faddism* to 1885, and so on. This work, which took a British team of lexicographers almost **40 years** to compile, makes a powerful impression. Here is an extract from one review:

How does one review a reference work which may well be the most important English dictionary compiled in the second half of the twentieth century (and a bit of the twenty-first)? The best strategy is probably to try to forget about this intimidating fact and treat the book as one would any other [Adamska-Sałaciak, 2010: 227].

We have already said that, according to our concept, the material or empirical basis for the chronologization of linguistic objects is **printed material.** This is not an absolutely necessary condition, since textual attestations may also appear in spoken form: in conversations, speeches, film dialogues, song lyrics and so on. Nonetheless, the most satisfactory results of chronologization, in terms of producing the most numerous and most certain dates, are given by analysis of graphical, that is printed, source materials.

3. Case study: work on the *National Polish Photocorpus* as an example of how to revolutionise dictionary creation

In order to create a database of two to three hundred thousand dictionary entries relatively quickly (in two to three years of work by a team of three to five persons) it will be necessary to undertake the actions described below.

3.1. Library queries. Building of a resource collection. The foundation of any extensive excerption effort is the identification of sources – determination of which textual resources will be subjected to mining. There are two routes that may be taken, depending on whether manual or elec-

tronic excerption is being considered. For both types of excerption, the sources analysed will date from the period 1901–2000.

3.1.1. Range of genres. To ensure coverage of as wide as possible a range of styles in twentieth-century vocabulary, excerption will be carried out from documents representing all possible genres (fiction, non-fiction, reportage, textbooks, etc.).

3.1.2. Range of authors. In selecting textual sources one must balance two opposing criteria, namely: a) as great as possible a choice of authors (to obtain adequate diversity of subjects and styles in the vocabulary used); and b) maximum focus on texts by authors whose idiolects are identified in preliminary analyses as being particularly rich in various lexical (derivational) forms – excerption will therefore be concentrated on authors such as Brandys, Dobraczyński, Dobrzyńska, Feldman, Głowiński, Gombrowicz, Irzykowski, Iwaszkiewicz, Kisielewski, Klemensiewicz, Konwicki, Kowalski, Kupiszewski, Michalkiewicz, Nowaczyński, Nowak, Nowakowski, Orzeszkowa, Pankowski, Parandowski, Pelc, Porębski, Słonimski, Srokowski, Szczepański, Wasilewski, Wiechecki, Zdziechowski, Żeromski, etc.

3.1.3. Number of sources. For manual excerption, it is assumed that approximately 4000–5000 books will be used as sources. With the project expected to last a total of around 36 months, manual excerption is planned to last for a period of 24 months (from the 7th to the 30th month of the project). It is expected that on average, approximately 20–40 multiword units (phrasemes) can be obtained from each book. The result of this reading will be units that are to be excerpted and presented according to the project principles; that is, primarily those which are absent from the first extensive twentieth-century orthographic dictionary, W. Kokowski's *Słownik ortograficzny języka polskiego*. In the case of electronic excerption, the set of sources will be the library objects contained in the *dLibra* system.

3.2. Physical acquisition of sources. Printed sources will be acquired in cooperation with the collections departments of large libraries (the National Library, Warsaw University Library, Warsaw Provincial Public Library, etc.), which have significant quantities of duplicate items for disposal. These libraries will, by oral request, supply these duplicates free of charge to outside parties. A project representative or employee selects the relevant titles and collects them from the library storeroom. In the selection of books that are not required by the library, there are no limitations in terms of quantity or weight. The person making the choice decides whether or not a particular book is useful. This solution eliminates the need to obtain specific approvals, shipping lists, etc.

3.3. Preselection of sources in *dLibra* for electronic excerption. Electronic sources will be taken from the database of the *dLibra* system. Be-

cause *dLibra* contains more than 3 million sources, some preselection of documents must take place. The first step is to identify texts published in the twentieth century. Although the date is explicitly given in the *dLibra* metadata, many individual records (in certain libraries) fail to conform to the expected pattern. The following date formats have been identified in *dLibra* metadata:

1884	20 January 2010	[post 1741]	22 II 1763
1920.03.27	1936.11.18	[ок. 1930]	[ante 1945]
1785–1819	1983–	1852 November	19th
[ca 1914]	no date	1940	12 III 1763
[1836]	27.08. February	1944 (Ausgabe Nr 1)	1850 ?
[ок. 1850]	[ante 1945]	2011 (digital edition)	7 IX 1762
datowanie 1613 r.	2–8 February	30.06.2008	2010
[XVIII/XIX w.]	09 March	2009 (orig. edition)	[192?]
1877_1877	1800/1900	no date	[post 1658]
1935–11–12	16 May 2008	1919–1939	1954–1964

This makes it necessary to apply standardisation, that is, to convert the dates to a uniform format, so as to enable as many documents as possible to be obtained with dates in the range 1901–2000. Next, a selection of appropriate sources must be made from the scanned documents. The *dLibra* system records document types such as *poster, legal document, album, article, magazine article, atlas, newsletter, brochure, magazine, magazine supplement, audio document, electronic document, manuscript, official document, document of community life, personnel document, musical work, leaflet, biweekly, dissertation, official journal, bookplate, photograph, graphical work, calendar,* etc. For example, it should be decided, based on this list, to include objects with the label *book*, and to exclude those with the label *photograph*. A similar issue arises in relation to the definition of the language of a document. The resources in *dLibra* may be labelled as representing a particular language, or with the label *multiple languages, undefined, no language context,* etc. Only documents labelled as *Polish* will be included in the analysis.

In the next step – following determination of which documents are to be subject to electronic excerption – it is necessary to assess which of the sources are suitable for further processing, and which are to be rejected due to their insufficient technical quality. For example, documents

scanned from microfilm are not suited for further processing. These include, among others, annual volumes of *Dziennik Poznański* (with a few exceptions) and *Kurier Poznański* (again with a few exceptions, such as the volume labelled 1936.01.01 R.31 no. 1, which is binarised – that is, presented in black-and-white form).

Next it must be verified whether the documents that have been scanned can be used for electronic excerption. Inadequate quality is found, for example, in issues of *Dziennik Białostocki* (e.g. 1928.03.28 R.6 no. 88), which, though binarised, is unsuitable for the process of electronic excerption. Generally speaking, the best OCR results are obtained from books rather than magazines.

3.4. Development of excerption algorithms. In all excerption projects it is of fundamental importance to define what action can be performed, within what scope and using what resources. In the case of the excerption of twentieth-century vocabulary, as has already been described, it is first necessary to form a collection of texts. In this project the texts will come from:

(i) the *dLibra* system; and

(ii) printed books and magazines.

What can be done to excerpt the words contained in the newspapers mentioned above? They may be read manually, but this is not feasible for a total number of several hundred thousand newspapers. Methods of automation must be sought. Five possible methods, labelled M1, M2, M3, M4 and M5, are described below.

In method M1, on a list of words obtained from digitised texts, we seek known words (for example, from a lexicographic source such as an orthographic dictionary). By this method we will not find any new words other than those that were previously known. We may, however, find real confirmation in the form of a quotation together with its location and chronologisation.

In method M2 we create an environment of artificial potentialisms by concatenating known elements (such as a list of prefixes and a list of words, taken for example from any Polish dictionary). In this method, again, we will not find units other than those whose elements (prefixes and root form) were known previously. It is therefore necessary to apply a different method.

In method M3, we seek words containing a particular consonant cluster. For example, knowing that in the Polish language, for instance in the source from method M1, the cluster *wst* occurs, we seek all other words with that cluster.

In method M4 we create possible artificial words (known as logatomes).

In method M5 we search a list of all words created from electronic documents in the list from methods M1 and M2, but with a wildcard character in place of one of the characters in the known word. This is based on the assumption of the existence of minimal pairs. This method will make it possible, to a certain extent, to find words from before the spelling reform of 1936, such as *Anglja* (from Angl*a).

3.5. Creation of a word list from electronic documents. All of the developed methods are applied to a word list created on the basis of the collection of electronic documents. This list is obtained from the textual layer produced by the OCR procedure.

3.6. Electronic excerption. In electronic excerption, the context surrounding the unit being searched for is cut out. The following criteria are used for selecting between quotations:

a) the graphical form of the text (a section is chosen with better scan quality, better resolution, better legibility, without physical defects such as folds or tears);

b) closeness to the start of each decade (from two alternative contexts for a given headword, a section from the older text is chosen; for example, for the headword *antyhitlerowski*, a quotation is taken from *Gazeta Bydgoska* dated 13 December 1931 rather than from *Orędownik Wrzesiński* of 3 August 1939);

c) the information content of the text fragment (quotations are preferred if they are easily assimilated, written in less complex language, taken from popular science works or teaching examples);

d) non-specialised nature of the text (there is a preference for texts addressed to a wide readership, such as daily newspapers);

e) spontaneous use, that is, instances that appear in lower case and without quotation marks.

3.7. Completion of database construction and online services. The quotations acquired in the manner described above will be made available to the general public via the website www.nfjp.pl. It will be possible to search the database both *a fronte* and *a tergo*. Searching will be possible based on the categories of (i) chronologisation and (ii) location.

4. Results

The parameters for the presentation of an excerption result are:

a) lemmatised dictionary entry;

b) documentary attestation (quotation);

c) attestation of location (title of publication);

3) chronological attestation (date of publication).

The headword is written in bold type, using the original spelling. The quotation must be presented in a freely readable form, using the original print size if possible.

An example of an excerption result appears below.

polihipowitaminoza

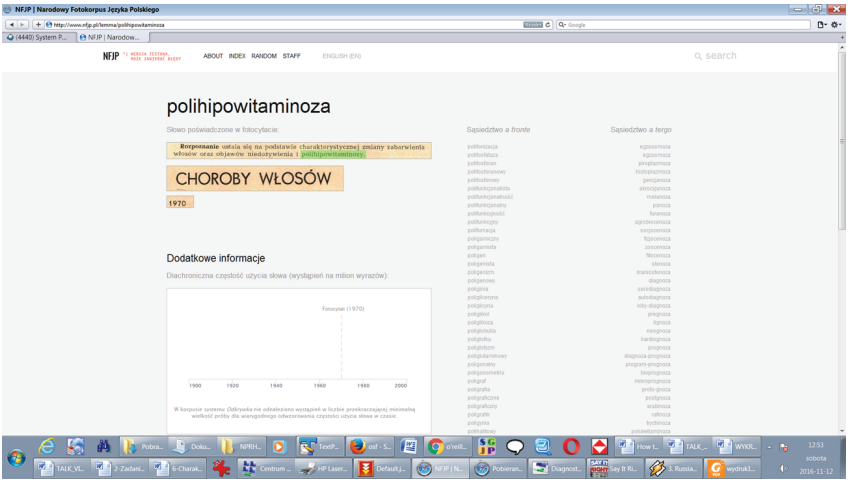
Rozpoznanie ustala się na podstawie charakterystycznej zmiany zabarwienia włosów oraz objawów niedożywienia i polihipowitaminozy.

'Diagnosis is determined by the typical changes of hair'colour and symptoms of malnutrition'

Bibliographic address:

Kostanecki, Wojciech 1970. Choroby włosów 'Illness of Hair', Warszawa : PZWL

The following image shows how this will be displayed on the www.nfjp.pl website:



5. Conclusions

The procedure described here will lead to the construction of a database that documents the use of lexical units. It will provide an important supplement to existing knowledge concerning the lexical resources of the Polish language. It is significant that for the first half of the twentieth century, particularly the two interwar decades, such records are extremely sparse, being limited to the systematic excerption efforts of two researchers, J. Wawrzyńczyk and P. Wierzchoń. The preliminary functionality of the

database can be tested at the website www.nfjp.pl. Under the present project it is planned to carry out excerption for approximately 250,000 to 300,000 units, more than the number of entries contained in any twentieth-century dictionary of Polish. The creation of such a database will also lead to progress in perfecting methods of electronic excerption, through the development and application of various excerption methods, such as the five described above.

The resulting lexical database will, firstly, be relatively large (the largest database of its type anywhere in the world); secondly, it will be suited to a variety of applications (morphological research, studies of borrowings, phraseological research, stylometric analyses, etc.); thirdly, it will enable viewing of the original context in photodocumentary form; and fourthly, it will contain tools to allow searching according to desired criteria.

Bibliography

Adamska-Sałaciak, A. rev.: *Kay, C.*, *Roberts, J.*, *Samuels, M.*, *Wotherspoon, I.* (eds). Historical Thesaurus of English. Oxford: Oxford University Press, 2009.International Journal of Lexicography, 2010. Issue 23/2. Pp. 227–233.

Bartmiński, J. (ed.) Współczesny język polski. Lublin, 2001.

Górny, M., *Wierzchoń, P.* Polish digital libraries as a philologist's tool. Based on 666 adjectives from the Digital Library of Wielkopolska. Poznań: IJ UAM, 2010.

Jadacka, H. System słowotwórczy polszczyzny (1945–2000). Warszawa: Wydawnictwo Naukowe PWN, 2001.

Kay, C., *Roberts, J.*, *Samuels M.*, *Wotherspoon, I.* (eds). Historical Thesaurus of English. Oxford: Oxford University Press, 2009.

Koerner, E.F.K., *Szwedek, A.* (eds.). Towards a History of Linguistics in Poland. From the Early Beginning to the End of the Twentieth Century. Amsterdam – Philadelphia: John Benjamins Publishing Company, 2001.

Matelski, D. Polityka Niemiec wobec polskich dóbr kultury w XX wieku. Toruń: Wydawnictwo Adam Marszałek, 2009.

Mazurek, C., *Stroiński, M.*, *Werla, M.*, *Węglarz, J.* (eds.) Polskie biblioteki cyfrowe 2008. Poznań: Ośrodek Wydawnictw Naukowych PAN, 2009.

Mazurek, C., *Stroiński, M.*, *Werla, M.*, *Węglarz, J.* Infrastruktura bibliotek cyfrowych w sieci PIONIER. Mazurek, C., Stroiński, M., Węglarz J. Polskie biblioteki cyfrowe 2008. 2009 (eds.). Pp. 9–13.

Piotrowski, T. Słowniki języka polskiego. Bartmiński J. (ed.) Współczesny język polski. Lublin, 2001. Pp. 601–618.

Piotrowski, T. Lexicography in Poland: From Early Beginnings – 1997. Koerner, E.F.K., Szwedek,A. (eds.). Towards a History of Linguistics in Poland. From the Early Beginning to the End of the Twentieth Century. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2001. Pp. 101–122.

Waszakowa, K. Przejawy internacjonalizacji w słowotwórstwie współczesnej polszczyzny. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego, 2005.

Wawrzyńczyk, J. (ed.) Czterdzieści lat minęło... nad „Słownikiem Doroszewskiego". Warszawa: TAKT, 2009.

Wawrzyńczyk, J. Nowe słownictwo polskie. Fikcje i fakty. Warszawa: Instytut Informacji Naukowej i Studiów Bibliologicznych, 1999.
Wawrzyńczyk, J. Słownik bibliograficzny języka polskiego. Wersja przedelektroniczna. Tom 1. Warszawa: Instytut Informacji Naukowej i Studiów Bibliologicznych, 2000.
Wawrzyńczyk, J. Autosuplement do Słownika warszawskiego. Poznań: Sorus, 2009.
Wawrzyńczyk, J. Słownictwo nowopolskie. Redatacje. Warszawa: Bel Studio, 2011.
Węglarz, J. Sterowanie w systemach typu kompleks operacji. Warszawa: PWN, 1981.
Wierzchoń, P. W poszukiwaniu czasowników nieznanych lingwist(k)om. Język. Komunikacja. Informacja2, 2007. Pp. 149–197.
Wierzchoń, P. Fotodokumentacja. Chronologizacja. Emendacja. Teoria i praktyka weryfikacji materiału leksykalnego w badaniach lingwistycznych. Poznań: IJ UAM, 2008a.
Wierzchoń, P. Jaskółki przejawów internacjonalizacji w słowotwórstwie współczesnej polszczyzny w materiałach z lat 1894–1984. Tylko sto przykładów. Łask: LEKSEM. 2008b.
Wierzchoń, P. ANTI. Poznań: IJ UAM, 2008c.
Wierzchoń, P. Kotuś. «Verba polona abscondita...» (w fotodokumentacji). Szkic lingwochronologizacyjny. Centuria pierwsza. Poznań: IJ UAM, 2008d.
Wierzchoń, P. Dlaczego fotodokumentacja? Dlaczego chronologizacja? Dlaczego emendacja? Instalacja gazowa, parking podziemny i „odległość niezerowa". Poznań: IJ UAM, 2009a.
Wierzchoń, P. Fotodokumentacja 3.0. Język. Komunikacja. Informacja 4, 2009b. Pp. 63–80.
Wierzchoń, P. Torując drogę teorii lingwochronologizacji. Investigationes Linguisticae XX. Pp. 105–186, 2010a.
Wierzchoń, P. Lingwochronografia na usługach słowotwórstwa gniazdowego. Kwartalnik Językoznawczy 1(2). Pp. 50–64, 2010b.
Wierzchoń, P. Depozytorium leksykalne języka polskiego. Nowe fotomateriały z lat 1901–2010. Vol. I. Warszawa: Bel Studio, 2010c.
Wierzchoń, P. Depozytorium leksykalne języka polskiego. Nowe fotomateriały z lat 1901–2010. Vol. II. Warszawa: Bel Studio, 2011a.
Wierzchoń, P. Depozytorium leksykalne języka polskiego. Nowe fotomateriały z lat 1901–2010. Vol. III. Warszawa: Bel Studio, 2011b.
Wierzchoń, P. Depozytorium leksykalne języka polskiego. Fotosuplement do Słownika warszawskiego.Vols. XI–XL. Warszawa: Bel Studio, 2014.