From Printed Materials to Electronic Demonstrative Dictionary – the Story of the National Photocorpus of Polish and its Korean and Vietnamese Descendants

Łukasz Borchmann, Daniel Dzienisiewicz, Piotr Wierzchoń

Institute of Linguistics
Adam Mickiewicz University
Poznań, Poland
E-mail: {borch, dzienis, wierzch} @amu.edu.pl

Abstract

The most popular form of lexicographic exemplification is plain-text transcript. Apart from the doubtless advantages of such a quotation method, it may be perceived as a kind of trade-off when considering readability, accessibility, simplicity, accuracy, and even the logistics of a documentation project. Another approach is to gather and present excerpts in the form in which they were originally published, that is, as the clippings from publications (this is referred to as *photodocumentation*).

The photodocumentary technique is a distinctive feature of both the National Photocorpus of Polish and its Korean and Vietnamese descendants. The main goal of the first of the above-mentioned projects was to describe around 250,000 lexical units, which would be enough to outperform all of the 20th-century dictionaries of Polish. Even more momentously, the process was entirely corpus-driven – that is, all of the principial lexicographic works preceding the project were intentionally ignored. As a result, the material contains largely the words of which linguists were unaware of or which were perceived as later neologisms under leading derivative models of Polish.

This article describes the projects from their early stages, namely the acquisition of printed materials, to the final level of development where an electronic lexicographic tool is made available to both amateur and professional users. Also described is the struggle to avoid unthinking imitation of p-lexicographic techniques. The methodology had to be adapted to meet modern web usability standards.

Keywords: e-lexicography; photodocumentation; corpus linguistics; computational linguistics; digitisation

1. Introduction

Lexicography, from a discipline built around traditional, deeply philological methods, has transformed into an interdisciplinary field involving both linguistics and computer science. This transformation is well reflected in many aspects of the National Photocorpus of Polish (NFJP) project and its Korean and Vietnamese descendants.

Three key ideas behind this lexicographic project are outlined in the following sections.

1.1 Photolexicography

Firstly, the project is based on photolexicography, a documented subdiscipline of applied linguistics in which every lexical unit is presented in exactly the same form as it appeared in print, along with its lexicographically relevant context (see Figure 1).

The method, which originated nearly a decade ago, is still progressing dynamically, not only contributing to the development of the basis for lexico-derivational models of 20th-century Polish, but also finding applications in a variety of new analyses, descriptions and glosses.

The advantage of the photodocumentary approach to quotation is that it prevents the risk of erroneous recreation or inaccurate recording of text, and, what is more, it presents maximally complete information, preserving both the textual contents and the original typographic layout (Małek, 2008; Wierzchoń, 2009).

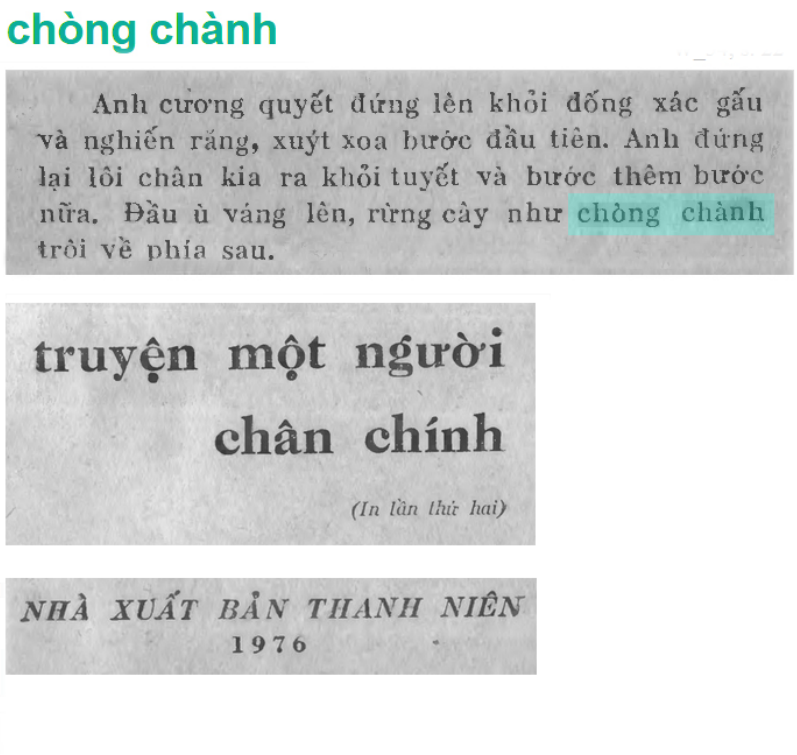


Figure 1: Vietnamese excerpt in the original form, that is, as a clipping from a publication (an example of *photodocumentation*)

1.2 Demonstrative dictionary

Оазис. *Образно.* ...Чтобы спасти *оазис* своей индивидуальности (А.И.Титаренко).

Обвод. По мысли Леонардо да Винчи, первый рисунок – это тень предмета, освещенного костром. Первобытный человек начинает рисовать, осваивая технику «*обвода*». Пещеры сохранили десятки таких примеров (В.Н.Дублянский).

Обгонять. Поменьше запретов – запреты уменьшают самостоятельность, ухудшают мышление, снижают ответственность. «Уверен – *обгоняй*» (Ю.Крелин).

Обезлюдеть. Дедушка рассказывал, что он отчетливо помнит два дня войны – начало и Победу. Он так описывал начало войны: «Из окна было видно море, легкий туман над ним. Солнце. Мы с ребятами проснулись рано, сидели за книжками. И вдруг за городом забухали пушки, зенитные разрывы в синеве неба, вой самолетов. Самым тягостным и тревожным было в тот день то, что Феодосия, казалось, враз *обезлюдела*. На улицах только военные» (ИМС, М.Полякова).

Figure 2: Extracted from *Словарь богатств русского языка*

Secondly, the NFJP project aims to create a demonstrative dictionary – a new type of work with its origins in Russian lexicography, as described in the 2003 work ***Словарь богатств русского языка*** (Figure 2; Kharchenko, 2003).

The authors of the original demonstrative dictionary aimed to present the wealth of the language and its curiosities of which people become unaware through everyday experience (Bobunova, 2013: 180). Aimed at the promotion of the lexical abundance of the Russian language, the project popularised, among others (Kharchenko, 2015):

- rare words discovered in texts and historical dictionaries, recorded with a view to reviving them;
- aphorisms that are not commonly known, mostly taken from the works of local writers from the 1970s, 1980s and 1990s;
- extracts from literary, popular-scientific and scientific texts where a given word was used in such a way that it deserved recognition and quotation;
- *biographemes* (*биографемы*), namely microdescriptions of family history and genealogical notes;
- attestations of the use of metaphors in the periods in which they were formed and when the motivational basis for formulating them was clear.

The above list does not exhaust the contents of the dictionary, but it enables us to comprehend the intentions of its authors of the enterprise. It also records idioms, sayings, proper names and lexical items used solely by particular authors.

There are numerous analogies between the premises of a photocorpus and the concept of a demonstrative dictionary, which lead us to consider NFJP a distinctive variety of the latter, referring to a related lexicographic tradition and a similar means of preservation and promotion of a national legacy.

Despite the fact that the two projects are closely related, one can distinguish methodological differences, which is evidenced by the fact that in its nature the demonstrative dictionary is a traditional work and the material contained in it is a result of decades of manual *gathering of words* (Kharchenko, 2015), as such an activity is described by (Małek, 2008).

1.3 Electronic lexicography

Thirdly, not only is NFJP a repository of lexical inventory, but it is also an e-lexicographic tool (for instance, involving such features as e.g. morphological tagging and searching with the use regular expressions – see Section 3.1).

Nowadays both the theory and practice of lexicography are deeply rooted in information science, which is reflected in the present work as well as in the NFJP project and methodology.

With the transformation of lexicography, the issue arose as to whether a theory setting a new direction for computational studies should be devised. Some claimed, however, that lexicographers should adhere to the concepts dating from the era of p-lexicography. A potential advantage of electronic dictionaries over traditional ones, as noted by (Nichols, 2010), is liberation from the limits set by the space taken by entries concerning their number and exemplifications as well as the length of the definition. Such limits are practically non-existent in the case of electronic dictionaries.

In the pre-electronic era the immediate elimination of errors was impossible – this difference is also indicated by (Nichols, 2010), who states that error correction can be performed online at any moment.

The above-mentioned possibilities can be recognised as reactions to problems of which traditional lexicographers are commonly aware. The advantage of e-lexicography is the

fact that a website constitutes a much more effective material than paper, due to its interactivity.

As a point of reference, one may consider a division of e-lexicographic tools into four categories (Tarp, 2011: 57–62):

1. digitised dictionaries, originally published in paper form;
2. dictionaries originally developed in a digital form, although with data structured as in traditional dictionaries – despite the more effective access (e.g. due to the headword search function) these are projects based on *utraditional models and concepts which have been taken over uncritically from the era of p-lexicography*;
3. tools with *dynamic contents and dynamically generated data*, crossing the borders of conventional lexicography, offering configurable functions enabling the dictionary to be adjusted to specific needs and expectations;
4. e-lexicographic tools, that are expected to be implemented in the future, which will enable one to combine the data from a previously prepared database with the data accessed online, so that it will be possible *de facto* to create and re-represent entries in real time.

One may familiarise oneself with real interactivity through two existing collections. These examples of projects from the third of the above categories are *Den Danske Ordbog* and the *Macmillan Dictionary and Thesaurus*.

Contrary to that which traditionally oriented scholars might claim, abandoning the idea of planning and developing a dictionary in its traditional form is a necessary step in order to access the broader perspective of contemporary lexicographic tools (Gouws, 2011).

Viewing online dictionaries as a search tool and abandoning the vision of a repository containing data or a conventional dictionary, allows their usability to be tested in a way which has been successfully applied to IT systems (see Heid, 2011).

1.4 Photographic quotation: a desirable practice or a foreign body in the world of e-lexicography?

The description contained in the preceding section may give the impression that a photographic quotation is in some ways incompatible with the idea of e-lexicography, and that NFJP might be considered an example of a project based on uncritically acquired models and concepts from the era of p-lexicography, as it was put by (Tarp, 2011).

The methods applied in the process of searching for textual attestations and edition of entries undoubtedly fit within the discipline of computational lexicography, and are far removed from the traditional conservative approach to lexicography (Piotrowski, 2001; Atkins & Zampolli, 1994; Boas, 2009). Is it not the case that a photographic quotation, being a digitised form of paper material, reintroduces old models and concepts into a world which has the aim of reforming them? A text presented in the form of raster graphics resembles the worst practices of website creation.

To avoid this situation, actions were taken to adapt the concept of photographic quotation developed for paper publications to the reality of modern lexicographic applications. While

photographic quotations were still demanded for each item, the contents of the exemplum were also required in the form of regular text. At the present stage of development of the project, this is text that is recognised automatically. In the future, manual verification will be made possible.

An exemplum obtained in such a way is used as the alternative text of a photographic quotation (for search engine robots and people with disabilities), but with the help of developed tools, phonetic transcription would be possible, for instance. In this way we attempt to combine the accuracy of documentation with the possibilities related to access to the content of the quotation.

Naturally, the above discussion does not exhaust the issue of the position of NFJP in the world of contemporary e-lexicography – this question, considered in more general terms, is addressed in the next section. The present study describes the projects from their early stages, namely the acquisition of printed materials, to the final level of development where an electronic lexicographic tool is made available to both amateur and professional users.

2. The process

Not to mention the problems of digitisation, difficulties abound even when the materials have already been scanned, analysed with OCR software and tokenised. Because of OCR errors, some kind of positive lookup is helpful in order to select promising lexical units for further analysis.

The following sections describe these difficulties, as well as the process of verification and editing of units by qualified annotators. Figure 3 is an illustration of the entire process of creating the NFJP resource described in this part of the article, and may be helpful in resolving any ambiguities.

2.1 Acquisition, preparation and preprocessing of the materials

At the current stage of the project's development, materials from in-house digitisation (referred to as the *non-electronic canon*) have been used in addition to materials from Polish digital libraries (the *electronic canon*). The non-electronic canon consists of approximately 4,000 books received free from non-electronic libraries which planned to recycle them, while 2,000 additional books from the electronic canon were selected to balance the corpora diachronically.

Information exchange at Polish digital libraries takes place using the OAI protocol. Most of the publications stored by *dLibra*[1] are from the pre-war period, up to 1939. The digital libraries also store various types of collections (printed matter, press cuttings, audiovisual materials). As a result, over a period of more than 10 years, a collection of over three million digitised library items has been built up. This material is described according to the Dublin Core scheme.

Unfortunately, the Polish digital library system does not offer normalised metadata, such as publication type or even year of publication, which are vital for many purposes. The

[1] A program used for the collection, editing and sharing of digital publications, developed at the *Poznań Supercomputing and Networking Centre*.

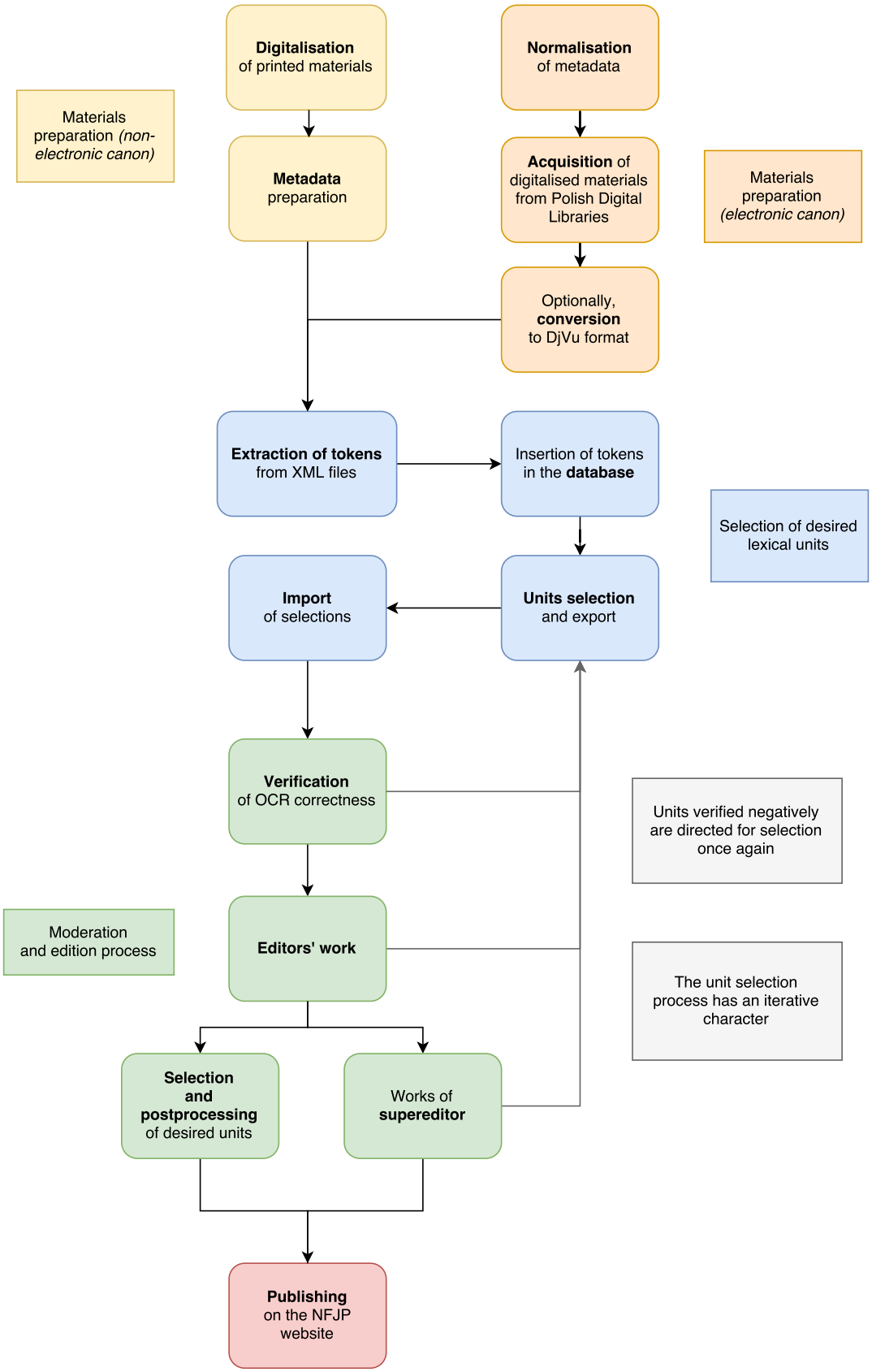


Figure 3: The process of creating the NFJP resource

structured data available via the OAI-PMH mechanism contain subject, type and date elements, but the practice of their use varies between and within libraries, so that automatic or semi-automatic normalisation had to be performed to convert this data to a form that would be easily usable by a computer program. Consider, for example, the following instances of text contained in the date field:

1884	rok obiegu 1940	1944 (Ausgabe Nr 1)
20 stycznia 2010	[ca 1914]	1850 ?
[post 1741]	b.d.	[ok. 1850]
[ok. 1930]	[192?]	[post 1658]
1920.03.27	1877	1800/1900
1936.11.18	22 II 1763	lata miedzywojenne
1785-1819	ante 1945	lata 30. XX w.
1983-	19w.	poczatek XIX w.
[XVIII/XIXw.]	12 III 1763	
mar-09	[1836]	
1852 November	27-lut-08	

Moreover, resources are available in different file types, so that within one digital library some publications may be published as multiple PDF files, and others as single or multiple DjVu files.

Before further processing, the materials obtained from these two heterogeneous sources were unified to single DjVu files, and for each of them XML files containing information about the text layer were created (with the use of the *djvutoxml* command from the *DjVuLibre* package). Years of publication from the electronic canon were normalised using a rule-based algorithm which selected the most pessimistic option, that is, the last year valid for a given textual date or period. Not only the date field was used, but also the title, which sometimes contains a more specific date (for example, there are cases where the date field contains a period, while there is a four-digit year within that period available in the title field).

2.2 Selection of lexical units for further processing

The content of an XML *word* tag was treated as a token, normalised, and inserted into a relational database with the structure presented in Figure 4 (names of tables and fields are self-explanatory). Obviously, not all of the unique tokens are correct Polish words (in fact, only around 10–15% are). To ensure low editing costs, because of OCR errors some kind of positive lookup needed to be used to select only promising lexical units for further analysis.

The first method that comes to mind is the use of dictionaries, and naturally this was attempted. However, the intention was to apply also a more sophisticated solution involving the generation of *verba possibilia*.

This term was coined to describe artificially created words on the basis of how morphological derivation works in a particular language. These few examples shed light on the method:

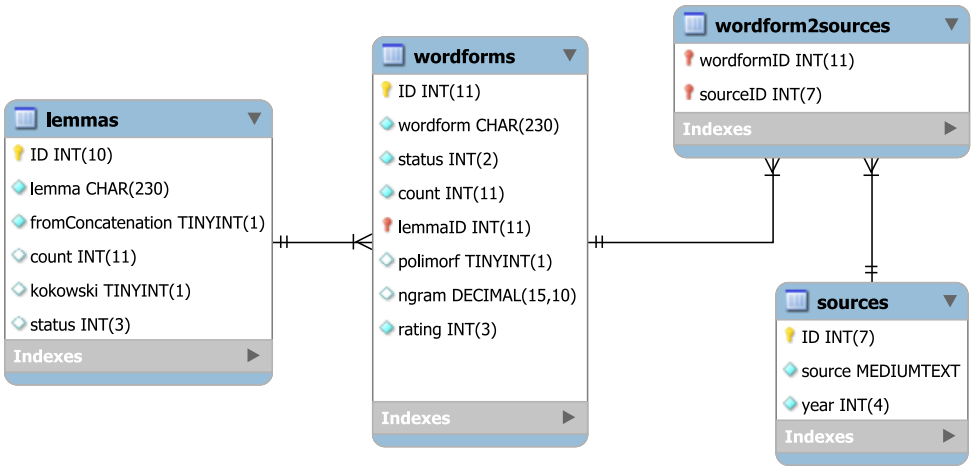


Figure 4: Schema of the database used in the process of selection of lexical units

- *naukowoczysty* 'scientifically clean' (concatenation of *naukowo* 'scientifically' and *czysty* 'clean');
- *panna-wdowa* 'spinster-widow';
- *samozaciemnienie* 'self-blackout' (concatenation of *samo* 'self' and *zaciemnienie* 'blackout').

One can also formulate rules to create unknown but probable words using the right-sided derivation, for example, using the equivalent of the English suffix *-zation/-sation* – Polish *-zacja*, Vietnamese *hóa* or Korean *화* (hwa):

- bình thường hóa 'normalisation'
- cách mạng hóa 'revolutionisation'
- chính thức hóa '*officialisation' (forms marked * probably do not exist within the English language, but the assumption that they will never be used in texts would be unreasonable)
- hoạt hóa '*activisation'
- hợp lý hóa 'organisation'
- 표준화 'standardisation'
- 세계화 'globalisation'
- 식민지화 'colonisation'

Many more unexpected findings can be obtained using two other methods applied within the NFJP project. The first of them is based on the assumption that unrecognised tokens that appear in a text in the context of known words are more likely to be correct Polish words than those which are never present in such a context. The second is the simple character-level n-gram word model (Jurafsky & Martin, 2000).

2.3 Verification and editing process

To verify the correctness of OCR and tokenisation, the panel shown in Figure 5 was prepared (the one shown was used during the preparation of the Great Photocorpus of

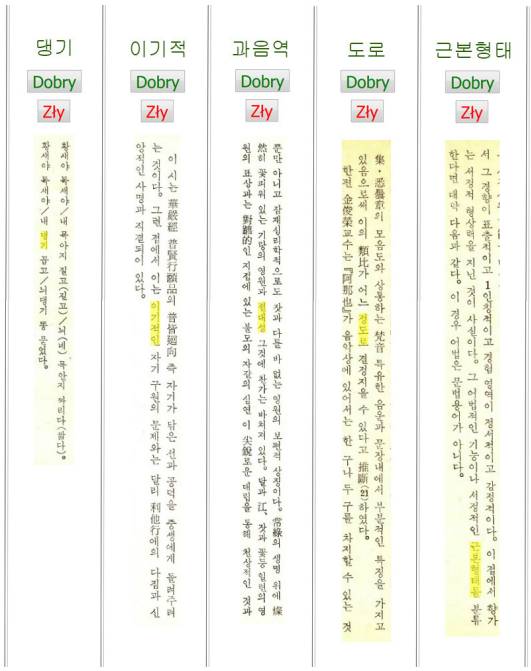


Figure 5: Initial verification of OCR for the purposes of the Great Photocorpus of Korean. The task of the reviewer was simply to check whether the highlighted word was equal to that recognised automatically

Figure 6: Editor's panel – part presenting the analysed unit

Korean, described more profoundly in Section 4; in case of other language variants it is analogous).

The approved units are then reviewed and annotated by editors with a strong linguistic background, who determine the lemma, the part of speech (in the case of phrases, instead of verb, for instance, verb phrase is presented as an option), and other grammatical categories (Figure 6). For the purposes of editing they are able to see the usage of the word in a broader context, up to the whole page.

During initial photodocumentation work, excerpts were cropped manually, as they were expected to meet certain rigorous conditions. Subsequently, as projects became more and more massive, steps were taken to make the cropping process fully automatic. Somewhat unexpectedly, the results of automatic methods proved to be indistinguishable from the manual ones, even without the use of machine-learning solutions. The currently utilised script uses heuristic methods based on recognised orthographic text (so as to take sentence beginnings and endings into account) and words' coordinates.

3. Functionality

The NFJP project is currently a fully functioning website, providing useful features for both amateur and professional users (see Figure 7 presenting entry structure). There are some new advanced features that will be released shortly; these will be discussed in a separate section below.

3.1 Publicly available

3.1.1 REGEX-based searching

The NFJP engine allows one to use Perl Compatible Regular Expressions while performing a search action. A systematic description of this formalism is not an aim of this work, thus we present only a few examples below.

The `$` character in REGEX syntax stands for an anchor to the end of the string. Thus the query `stylowy$` 'stylish, in style' would return results such as *ponadstylowy* 'abovestylish', *neostylowy* 'neostylish' and *emocjonalno-stylowy* 'emotionally-stylish'. Similarly, the `^` character matches the start of the string to which the regex pattern is applied; thus the query `^pseudo` would return such words as *pseudozdrajca* 'pseudotraitor', *pseudowynalazca* 'pseudoinventor', *pseudoszwabacha* 'pseudoschwabacher (a specific blackletter typeface)'.

A slightly more advanced example of a regular expression is `^.{4}$`, which returns words consisting of exactly four characters.

For more advanced examples of regular expressions usage see Friedl (2006), Good (2004) and Stubblebine (2003).

3.1.2 Search operators

Modern search engines provide a feature allowing one to make search results more precise using so-called search operators. A similar solution is implemented in the National Photocorpus.

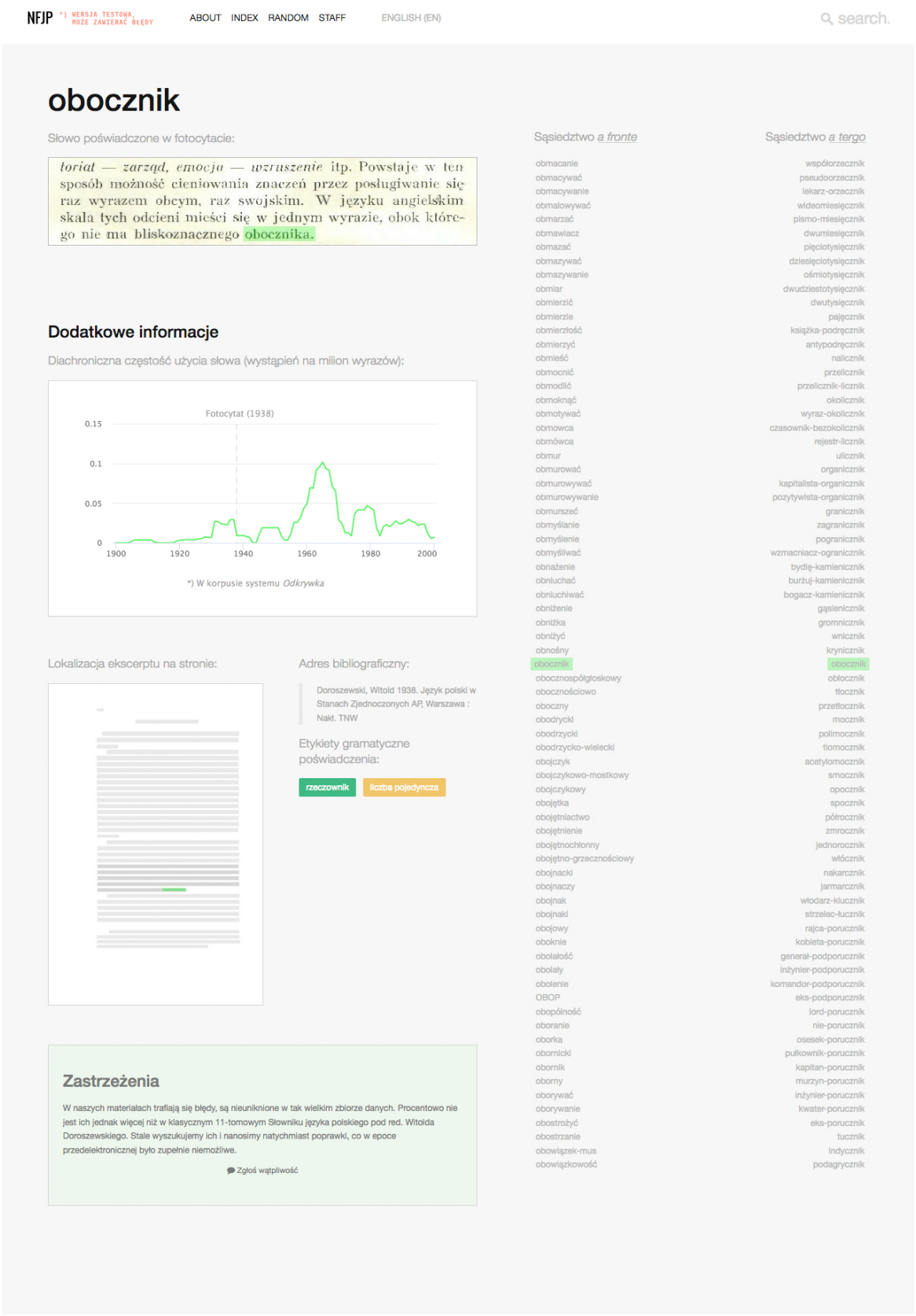


Figure 7: View of the entry for the word *obocznik*

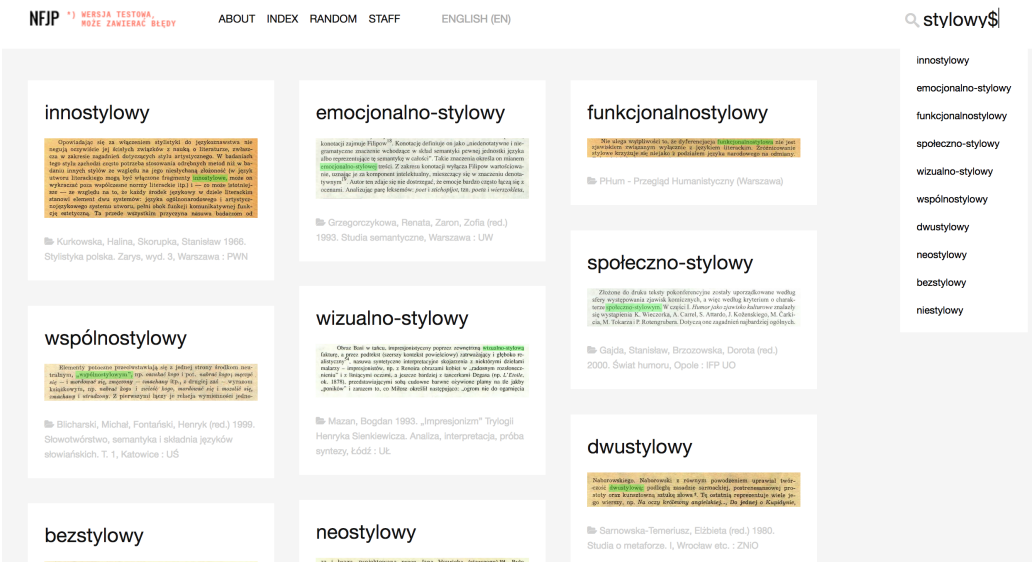


Figure 8: Results obtained with the use of a regular expression

Part of speech. Using the *pos* operator one can return results matching only the selected part of speech. Available values are: *verb, part, num, particle, pred, prep, adj, adv, subst, conj, interj, ppron, other*. For example, adding `pos:adj` to a query will cause it to return only adjectives.

Number. The string `number:pl` in a query will restrict the results to plurals only. Other available values are *sg, pt* (pluralia tantum) and *du* (dual).

Source. The string `source:IJ_698` in the search input will return only words found in the book *Encyklopedia techniki. Przemysł spożywczy* (Banecki et al., 1978), because *IJ_698* is its ID within the system.

Reflexive form. For the purposes of binary features, the feature operator was introduced. At present it allows one to restrict the results to reflexive verbs using `feature:reflexivum`.

Multiple search operators can be used in one query and they can be combined with regular expressions. For example `^s source:IJ_2788` will return words beginning with the letter *s* from the source with the selected ID.

3.1.3 *A fronte* and *a tergo* neighbourhood

On the details page of each entry, *a fronte* and *a tergo* neighbourhoods are presented. For example, for the entry *ślimaczenie sie* such a neighbourhood is:

ślicznіuchny	próżniaczenie
ślicznіutki	półmajaczenie
ślicznіutko	żydłaczenie
ślicznotka	rozkułaczenie
ślimaczenie sie	**(sie) ślimaczenie**
ślimaczo	przysmaczenie
ślimakowato	re-tłumaczenie
ślimakowo	przetłumaczenie
ślimakowo-wirnikowy	idiotłumaczenie

On the NFJP website 36 words above and below the displayed unit are visible (Figure 7), which is useful particularly in a research regarding word formation and inflection (Grzegorczykowa & Puzynina, 1973; Obrebska-Jabłońska et al., 1968).

3.1.4 Other features and materials

For each of the words relative usage frequency is shown, within the period 1900–2000 (count per million words in publications from each year). See Figure 9.

The website also contains materials in five languages (Polish, German, English, Russian and Japanese) describing the purpose of the project, its methodology and the significance of the results, as well as information regarding other projects focused on Polish vocabulary undertaken prior to NFJP, a bibliography, and a library containing information about all of the publications describing NFJP materials.

3.2 Case studies

3.2.1 Lexical inventions of Adolf Nowaczyński

The authors of the work *Archikastrat, emancypaństwo i krytykretyni...* analysed the linguistic creativity of Adolf Nowaczyński, a Polish writer, poet, playwright, critic, and social and political activist (Dzienisiewicz et al., 2017).

In the course of the analysis the authors distinguished five categories: words which had been commonly used before they first appeared in Nowaczyński's works (A), words which had occurred several times before they first appeared in Nowaczyński's works (B), words with single or several occurrences after they first appeared in Nowaczyński's works (C), words whose use might have originated within Nowaczyński's idiolect (D), and words discovered solely in Nowaczyński's writings (E).

To perform analyses of this type, one may utilise two functions available in NFJP: the diachronic frequency of a word, and the search operator source:, allowing one to select all of the units recorded for the first time in a given publication.

One of the publications included in the NFJP canon is *Góry z piasku* by Adolf Nowaczyński, where such units as *afiszowość, aluzjonizm, junaczość, omłacanie, katastrefa, powstydzenie, wyklecić, proteuszowo, regencki, renomista, zniewieścialec, lubownictwo, nieobmieciony, nawałesać sie, mieszczuszek, nieprzyłaczony, nierozpowity, nierozjatrzanie,*

niedźwigajacy, niekabłakowaty, nieświatowość, oblagowywanie and *złotorunny* were discovered.

Most of the presented words are especially interesting in terms of their word-formative features, e.g. *zniewiescialec* (a personal noun denoting 'an effeminate man'), *złotorunny* (an adjective derived from the phrase 'Golden Fleece'), *powstydzenie* (an unusual form of the word 'ashamedness' with the prefix *po-*; the common Polish form is *zawstydzenie*), *mieszczuszek* ('a little city slicker'; an original example of the use of the diminutive suffix *-ek*).

Some of the above units were included in the categories devised by the authors; however, some of them were not recorded by them, although they meet the criteria for category E, that is, words discovered only in Nowaczyński's writings. The corpus of the Discovermat system (which served as a point of reference for the authors) returns one result for the query *junaczość* from an article by Nowaczyński published in *Nowy Przeglad Literatury i Sztuki*.

3.2.2 NRF and RFN

In the period of the Polish People's Republic two names were used to denote Western Germany, namely, *Niemiecka Republika Federalna* (NRF) and *Republika Federalna Niemiec* (RFN). Both abbreviations are included in NFJP, thus their diachronic frequency of occurrence in texts can be traced (Figure 9; Dzienisiewicz, 2017).

3.3 Russian and Soviet lexical borrowings

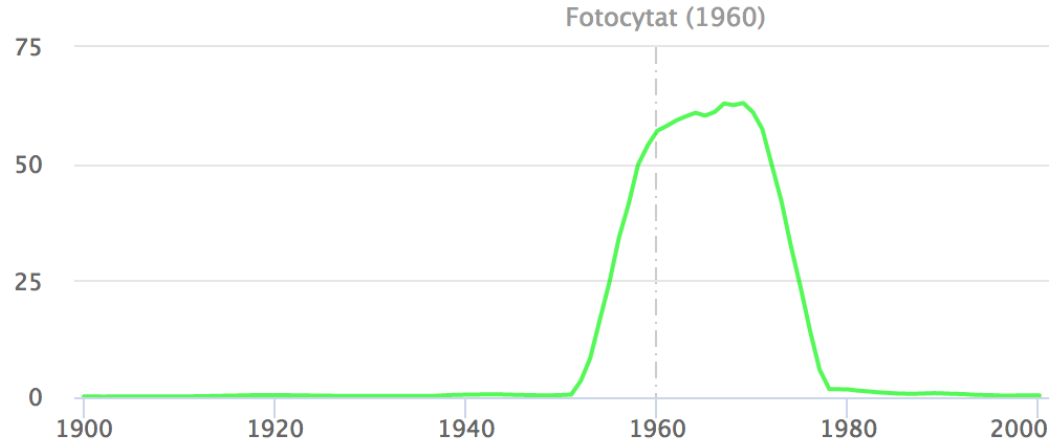
The list of publications available on the NFJP website enables one to distinguish several groups of sources which might include Russian and Soviet lexical borrowings, that is (Wawrzyńczyk, 2014):

- translations of Russian literary works (Chekhov, Dostoyevsky, Gogol, Lermontov, Pushkin, Solzhenitsyn, Tolstoy);
- translations of journalistic writings, diaries, letters and scholarly texts of, among others, Byelinsky, Herzen, Dostoyevsky, Zinovyev, Likhachov;
- diaries and correspondence of the Polish people who were sent to Russia and the USSR;
- works by Polish authors who lived in the Russian Partition.

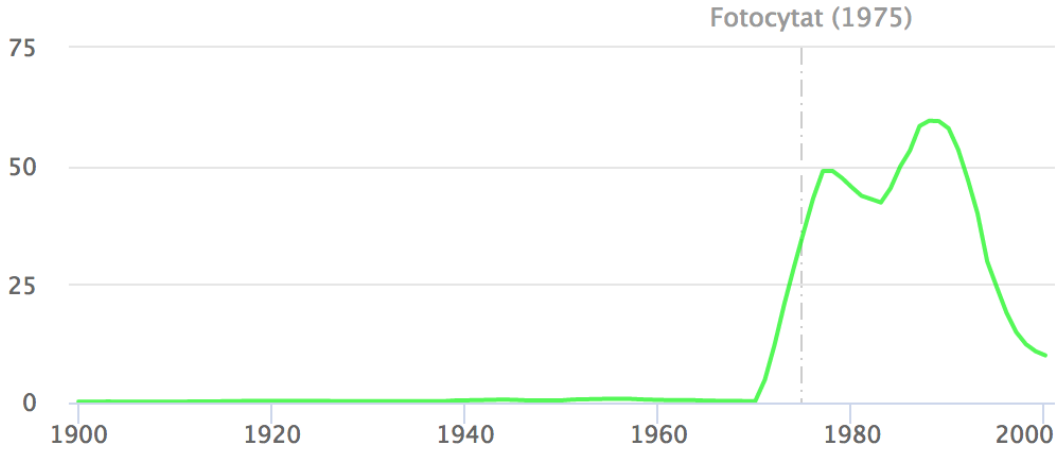
Using the source: operator one can obtain a list of words recorded for the first time in the above works. Even a cursory overview of the units brings to light some which might be of interest to scholars specialising in Russian borrowings, as it includes the following words: *niepuszkinowski, grażdański, sowchozowy, sofista-słowianofil, półimperiał*.

Even more interesting cases of words can be found in Pushkin's works (included in NFJP):

- the word *niedaleczko* discovered in the saying *Rzekłbym słóweczko, lecz wilk niedaleczko* (***Сказал бы словечко, да волк недалечко***, 'walls have ears');



— Chyba pomyłka, Panie Inżynierze — odpowiadał Doktor Piotrowski z maleńką nutką niepokoju w głosie. — To było specjalnie dla mnie sprowadzane z NRF-u. Zygmuś mi się zresztą z tym szybko uwinął. Choć co prawda słono extra kosztowało, ale miałem wóz na chodzie raz dwa. Ten chłopak po prostu cuda umie robić! Nie pytam już nawet jak...



takich zapisów nie spotyka się wcale. Oczywistą sprawą jest także to, że końcówki przypadkowe wyrażają skrótowce w wyższym stopniu zleksykalizowane, częściej używane lub należące do warstwy słownictwa potocznego, np. *piłkarze LZS-ów*, *RFN-u*, ale *ZSRR*.

Figure 9: Photographic quotations and diachronic frequency of occurrence of NRF (upper graph) and RFN (lower graph)

- the word *dych*, which appeared in the expression *ani słychu, ani dychu* (***Ни слуху ни духу***, 'there has been no news of somebody or something').

With the use of the described method a large-scale analysis of Russian borrowings can be conducted on the materials contained in NFJP.

3.4 Features to be released shortly

3.4.1 Morpheme segmentation

The automation of morpheme segmentation is not a trivial task and can be performed in various ways. Considering the fact that there are no large sets of annotated data for many languages and that creating them requires a huge amount of work, solutions based on unsupervised machine learning (Creutz & Lagus, 2007, 2005; Goldsmith, 2001) and minimally supervised machine learning techniques are popular. In the latter case models are learned from a small number of segmented words and a large number of unsegmented words (Ruokolainen et al., 2016). Fortunately, there are publications for Polish that make supervised machine learning techniques applicable without the need for additional annotating efforts, so that we can easily compare the performance of both approaches.

For the purposes of supervised machine learning two volumes of *The Dictionary of Derivational Nests of Modern Polish* were used (Jadacka & Bondkowska, 2002; Vogelgesang, 2001) with a total of 50,000 words. They required a pre-processing stage before performing supervised learning, because the format used was not segmented orthographic text. The only methodological difference between source segmentation and the one used in the described set is the abandoning of the null morpheme concept, which has no rational motivation in morpheme segmentation (nor in linguistics in general, cf. Mańczak, 1996: 11).

During the work the above set was split into random training and test subsets to perform cross-validation. The rule-based model was used as a baseline for machine learning techniques. It is similar to the one described by Yang (2007) but is simpler and based on a predefined list of morphemes.

In terms of supervised machine learning techniques, the problem of morpheme segmentation can be treated as a problem of binary classification, that is whether the morpheme boundary should or should not be placed between certain letters in a word (this approach is similar to the one described by Neubig et al., 2011 for Japanese). In order to determine the best classifier for this purpose, various methods available in the *scikit-learn* Python library were tested (Pedregosa et al., 2011). For each of the classifiers Confusion matrix was computed as well as other evaluation metrics, such as Accuracy, F1 score and Matthews correlation coefficient (MCC).

The optimal set of features seems to be similar to some of the features proposed for Arabic by (Monroe et al., 2014). In the case of the Polish language it consists of:

- a five-character window around the analysed character boundary;
- character n-grams made from the current character and up to the next four characters;

- character n-grams made from the current character and up to the previous four characters.

From the methods available within *scikit-learn*, only Decision Trees offers comparable results. Although the results of Decision Trees are weaker than those obtained using a linear Support Vector Classifier, its moderate effectiveness encourages us to check the results of combining both Decision Trees and SVC, using for instance a Voting Classifier. The idea is to combine different machine learning classifiers and use the average of the predicted probabilities offered by each of the combined methods. The method described, however, does not produce significantly better results.

A different approach to morpheme segmentation is to use a Conditional Random Fields statistical sequence modelling framework (Tseng et al., 2005). The problem is basically to predict a vector $y = \{y_0, y_1, \ldots, y_T\}$ of variables for a feature vector x. It can be solved by learning an independent per-position classifier that maps $x \mapsto y_s$ for each s, as was done in the above section, ignoring the sequential aspect of the data. By contrast, Conditional Random Fields refers to neighbouring samples and predicts a sequence of labels for a sequence of input sample (Sutton & McCallum, 2012).

For the purposes of this work, CRFsuite was used (Okazaki, 2007). This offers various training methods (such as Limited-memory BFGS, Orthant-Wise Limited-memory Quasi-Newton, Stochastic Gradient Descent, Averaged Perceptro, Passive Aggressive, Adaptive Regularization Of Weight Vector) and simple TSV input format.

The final CRF-based solution performed as efficiently as the best SVM-based solution in terms of evaluation metrics, even though it seems to outperform it when examining the results. It uses the Passive Aggressive training method (Crammer et al., 2006) and the following features (let c[t] be the current character in a word):

- a five-character window around the analysed character boundary (c[t- 2]|c[t-1]|c[t]|c[t+1]|c[t+2]);
- character n-grams made from the current character and up to four following characters (e.g. c[t]|c[t+1] for a bigram);
- character n-grams made from the current character and up to four previous characters (e.g. c[t-2]| c[t-1]|c[t] for a trigram);
- every single character within the word identified as e.g. c[t-4];
- c[t-2]|c[t-1] and c[t+1]|c[t+2];
- c[t-2]|c[t] and c[t]|c[t+2]=n|e.

Moreover, a family of methods for unsupervised learning of morphological segmentation was tested (e.g. one utilizing probabilistic generative models), as well as semi- (minimally) supervised machine learning (including a model trained on the full *National Corpus of Polish* skipping compounds with a random probability, this being expected to speed up the training considerably with only a minor loss in model performance; cf. Virpioja et al., 2013).

None of these attempts, however, resulted in a level of performance comparable to those obtained using the final SVM- and CRF-based models.

The features proposed in the literature for unrelated languages such as Chinese and Japanese are applicable to Polish with only minor modifications. The fact that the performance limit for three conceptually different methods stands at a similar level suggests that it is either a limit of machine learning methods (at least at this level of advancement) or a limit of training on the data set described in this paper. Observation of incorrect classifications reveals that they are sometimes related to the idea behind the *Dictionary of Derivational Nests of Modern Polish*, where some derivatives are presented without inherited morphological structure. This supports the second hypothesis.

Future work will focus on developing better training sets and on testing deep learning methods, as well as other ensemble combinations. Independently of this, the solution described in the present chapter is production-ready, and will be released shortly on the NFJP website.

3.5 Phonetic and phonematic transcription

Maria Steffen-Batóg proposed mechanisms of phonetic and phonematic transcription for Polish, based solely on the character context of a particular letter. The algorithm assumes iterative reading of input orthographic text (character by character) and matching of appropriate left and right context definitions from the tables of rules created by Steffen-Batogowa (1975) and Steffen-Batóg & Nowakowski (1997). In each of the tables the first row contains a formal definition of the right context, and the first column a definition of the left context. The proper transcription can be found at the intersection of the matching definitions.

The proposed formal definitions of left and right context (ca. 500 unique descriptions and many more combinations) were implemented using regular expressions. The correctness of the algorithm is currently being checked on the vast material of NFJP, and required fixes are continuously applied.

3.6 The formal definition of neologism

Matyka (2010) formulated three questions regarding neologisms:

- How can one objectively check whether a word is a new one?
- How one can determine its age?
- When should a lexicographer assume that a neologism is old enough to place it in his dictionary?

Answers to these and similar questions should consider that a word may be widespread within one group, but completely unknown within another.

For this purpose the Herfindahl–Hirschman Index was adapted. This is a measure of the size of companies in relation to their industry, widely applied in competition law as an indicator of the degree of competition (Calkins, 1983). It is expressed as the sum of squares of the shares:

$$HHI = \sum_{i=1}^{N} s_i^2$$

The HHI is the same as Simpson's index (Magurran, 1988: 39–40) used in ecology to measure the concentration of individuals classified into types (the two indices were proposed independently for analogous purposes). The HHI has also been used outside these fields, for instance to quantify level of political competition (Davidson et al., 2008).

In our case it reflects the concentration or dispersion of word usage among sources. A high value means that there are only a few sources to which the majority of word usage cases belong. The smaller the value, the greater the dispersion of the word among sources from a given year.

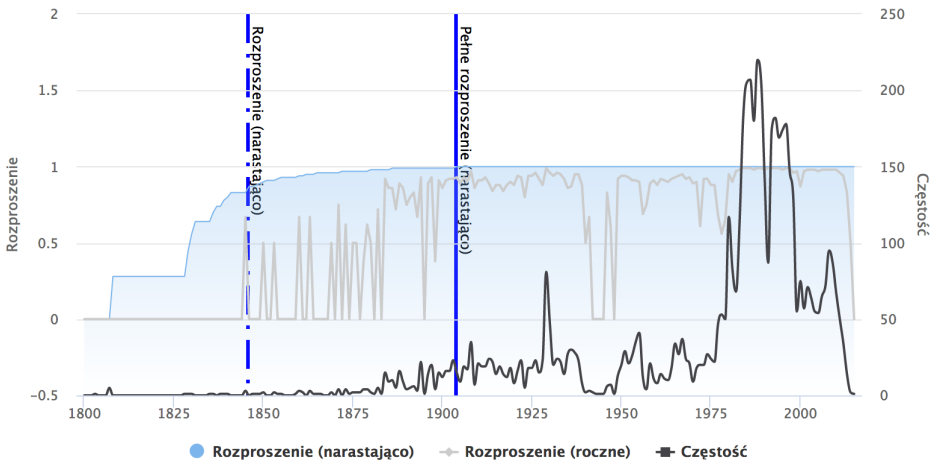


Figure 10: Word usage dispersion

In Figure 10 vertical lines denote some key moments, namely when HHI for the first time took a value smaller than 0.2 (interpreted in law and economics as unconcentrated industry) and the value 0 (highly competitive industry).

4. Discussion and perspectives

In the course of the development of NFJP, other e-lexicographic projects were derived from the original undertaking, namely the Great Photocorpus of 20th-Century Vietnamese and the Great Photocorpus of Korean. Created with the use of techniques developed while working on NFJP, the new enterprises provide us with some insights about the application of the original methodology to languages that are genetically unrelated to Polish.

Because in Vietnamese spaces are used not only to separate words, but also syllables (which may be words in themselves), from the perspective of photodocumentation procedures and software developed originally for Indo-European languages, such as Polish, an attempt to process Vietnamese words resembles in some way a multi-word expression analysis. Indeed, what we have done is treat Vietnamese words exactly as Polish multiword units within our system. The main difference relates to the above-mentioned problem; however, it is common to almost every natural language processing task involving Vietnamese, and thus has well-established solutions proposed in the literature. We decided to rely on the vnTokenizer, utilising the hybrid approach to word segmentation (Hông Phuong et al., 2008).

Figure 11: Newspaper from the 1970s with headlines written horizontally and article content vertically

In the Korean project a new problem arises, related solely to the automatic excerpt generation mechanism: text can be written either horizontally from left to right or vertically from top to bottom. What is more, both writing styles may be used on the same page, as shown in Figure 11.

The rest of the workflow, for both Korean and Vietnamese, remains almost entirely the same.

Despite the advancement of some features presented in this paper, plans are much more ambitious – for example, we intend to use methods generally not applied in the humanities, such as *word2vec* software, which can be used to determine semantic and syntactic relations between words (Mikolov et al., 2013c,a,b). These can be used in many ways – from simple visualisation of semantics to finding diachronic synonyms of a word and tracking changes of word meanings.

The future is near and will be even more e-.

5. Acknowledgements

Work supported by the Polish Ministry of Science and Higher Education under the National Programme for Development of the Humanities, 0014/NPRH3/H11/82/2014, *Narodowy Fotokorpus Jezyka Polskiego. Fotodokumentacja słownictwa XX w.* (National Photocorpus of the Polish Language).

6. References

Atkins, B. & Zampolli, A. (1994). *Computational approaches to the lexicon.* Oxford University Press.

Boas, H.C. (2009). *Multilingual FrameNets in Computational Lexicography: Methods and Applications.* Trends in Linguistics. Studies and Monographs 200. Mouton de Gruyter, 1 edition.

Bobunova, M. (2013). ***Русская лексикография** XXI века. **Учебное пособие***. Москва: **Флинта**.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S. & Singer, Y. (2006). Online Passive-Aggressive Algorithms. *The Journal of Machine Learning Research*, 7, pp. 551–585.

Creutz, M. & Lagus, K. (2005). Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology.

Creutz, M. & Lagus, K. (2007). Unsupervised models for Morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1).

Dzienisiewicz, D. (2017). Na krzyż: NRF vs. RFN. http://re-research.pl/pl/post/2017-01-30-60105-na-krzyz-nrf-vs-rfn.html.

Dzienisiewicz, D., Graliński, F. & Wierzchoń, P. (2017). Archikastrat, emancypaństwo i krytykretyni – głos lingwochronologizatorów w sprawie kreatywności jezykowej Adolfa Nowaczyńskiego. In *Kreatywność jezykowa w przestrzeni publicznej*. In print.

Friedl, J. (2006). *Mastering Regular Expressions: Understand Your Data and Be More Productive*. O'Reilly Media.

Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Comput. Linguist.*, 27(2), pp. 153–198. URL http://dx.doi.org/10.1162/089120101750300490.

Good, N. (2004). *Regular Expression Recipes: A Problem-Solution Approach*. Apresspod Series. Apress. URL https://books.google.pl/books?id=3ttQAAAAMAAJ.

Gouws, R. (2011). Learning, Unlearning and Innovation in the Planning of Electronic Dictionaries. In *E-Lexicography: The Internet, Digital Initiatives and Lexicography*. London: Bloomsbury Publishing, pp. 17–29.

Grzegorczykowa, R. & Puzynina, J. (1973). *Indeks a tergo do Słownika jezyka polskiego pod redakcja Witolda Doroszewskiego*. PWN.

Heid, U. (2011). Electronic Dictionaries as Tools: Toward an Assessment of Usability. In *E-Lexicography: The Internet, Digital Initiatives and Lexicography*. London: Bloomsbury Publishing, pp. 287–304.

Hông Phuong, L.ê., Thi Minh Huyên, N., Roussanaly, A. & Vinh, H.T. (2008). *A Hybrid Approach to Word Segmentation of Vietnamese Texts*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 240–249. URL http://dx.doi.org/10.1007/978-3-540-88282-4_23.

Jadacka, H. & Bondkowska, M. (2002). *Gniazda odrzeczownikowe*, volume 2 of *Słownik gniazd słowotwórczych współczesnego jezyka ogólnopolskiego*. Universitas.

Jurafsky, D. & Martin, J.H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1st edition.

Kharchenko, V. (2003). *Словарь богатств русского языка: редкие слова, метафоры, афоризмы, цитаты, биографемы*. Number t. 1-2 in Словарь богатств русского языка: редкие слова, метафоры, афоризмы, цитаты, биографемы. Изд-во Белгородского государственного университета.

Kharchenko, V. (2015). О демонстративном словаре русского языка. *Лексикография и коммуникация - 2015 : материалы I междунар. науч. конф.*, pp. 79–88.

Matyka, A. (2010). *Słowa – kładki, na których spotykaja sie ludzie różnych światów*, chapter O pojeciu neologizmu w jezykoznawstwie. Warszawa: Wydział Polonistyki UW, pp. 99–109.

Małek, E. (2008). *Ku fotoleksykografii*. Łódź: Instytut Rusycystyki Uniwersytetu Łódzkiego.

Mańczak, W. (1996). *Problemy jezykoznawstwa ogólnego*. Zakład narodowy im. Ossolińskich.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781. URL http://arxiv.org/abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K.Q. Weinberger (eds.) *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., pp. 3111–3119. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Mikolov, T., Yih, S.W.t. & Zweig, G. (2013c). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics, pp. 746–751. URL https://www.microsoft.com/en-us/research/publication/linguistic-regularities-in-continuous-space-word-representations/.

Monroe, W., Green, S. & Manning, C.D. (2014). Word Segmentation of Informal Arabic with Domain Adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 206–211. URL http://www.aclweb.org/anthology/P14-2034.

Neubig, G., Nakata, Y. & Mori, S. (2011). Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 529–533. URL http://dl.acm.org/citation.cfm?id=2002736.2002841.

Nichols, W. (2010). *English Learners' Dictionaries at the DSNA 2009*, chapter I've heard so much about you: Introducing the native-speaker lexicographer to the learner's dictionary. Tel Aviv: K Dictionaries, pp. 29–43.

Obrebska-Jabłońska, A., Dulewicz, I., Grek-Pabisowa, I. & I., M. (1968). *Indeks a tergo do Materiałów do słownika jezyka staroruskiego I.I. Srezniewskiego*. Państwowe Wydawnictwo Naukowe.

Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs). http://www.chokkan.org/software/crfsuite/.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Piotrowski, T. (2001). *Zrozumieć leksykografie*. Wydawnictwo Naukowe PWN.

Ruokolainen, T., Kohonen, O., Sirts, K., Grönroos, S.A., Kurimo, M. & Virpioja, S. (2016). A Comparative Study of Minimally Supervised Morphological Segmentation. *Computational Linguistics*, 42(1), pp. 91–120.

Steffen-Batogowa, M. (1975). *Automatyzacja transkrypcji fonematycznej tekstow polskich*. Warszawa: PWN.

Steffen-Batóg, M. & Nowakowski, P. (1997). An algorithm for phonetic transcription of orthographic texts in Polish. In *Studies in phonetic algorithms*. Poznań: Soros, pp. 581–602.

Stubblebine, T. (2003). *Regular Expression Pocket Reference*. Sebastopol, CA, USA: O'Reilly & Associates, Inc., 1 edition.

Sutton, C. & McCallum, A. (2012). An Introduction to Conditional Random Fields. *Foundations and Trends® in Machine Learning*, 4(4), pp. 267–373. URL http://dx.doi.org/10.1561/2200000013.

Tarp, S. (2011). Lexicographical and Other e-Tools for Consultation Purposes: Towards the Individualization of Needs Satisfaction. In *E-Lexicography: The Internet, Digital Initiatives and Lexicography*. London: Bloomsbury Publishing, pp. 54–70.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D. & Manning, C. (2005). A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. In *Fourth SIGHAN Workshop on Chinese Language Processing*. pp. 168–171.

Virpioja, S., Smit, P., Grönroos, S.A. & Kurimo, M. (2013). Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline. Technical Report 25/2013 in Aalto University publication series SCIENCE + TECHNOLOGY, Department of Signal Processing and Acoustics, Aalto University, Helsinki, Finland. URL https://aaltodoc.aalto.fi/handle/123456789/11836.

Vogelgesang, T. (2001). *Gniazda odprzymiotnikowe*, volume 1 of *Słownik gniazd słowotwórczych współczesnego jezyka ogólnopolskiego*. Universitas.

Wawrzyńczyk, J. (2014). *Jezyk, literatura i kultura rosyjska na stronie www.nfjp.pl*. Warszawa: Mila Hoshi.

Wierzchoń, P. (2009). Fotodokumentacja 3.0. *Jezyk. Komunikacja. Informacja*.