New Information Extracting and Analysis Methodology for the Terminology Research Purposes: the Field of Biology

Gints Jasmonts [1], Silga Sviķe [2] and Karina Šķirmante [3]

[1] Ventspils University of Applied Sciences 1, Inženieru Street 101, Ventspils, LV-3601, Latvia
[2] Ventspils University of Applied Sciences 1, Inženieru Street 101, Ventspils, LV-3601, Latvia
[3] Ventspils University of Applied Sciences 1, Inženieru Street 101, Ventspils, LV-3601, Latvia

Abstract

This study is part of a larger research developing a database of organism names for preserving linguistic diversity and for terminological studies. This article addresses the issue of retrieving the names of organisms from digitized resources of specialized literature for subsequent collection in the new database. The study used image processing algorithms, Optical Character Recognition (OCR) and language recognition algorithms, altogether creating a methodology that reflects the possibility of using data retrieval for the terms of the field of biology in Latvian and other local languages – the names of organisms along with their scientific or Latin name. Further on, these retrieved data are scrutinized and prepared for addition in the developed database.

Keywords

Information extracting, organism names, database, digitized resources, OCR methods

1. Introduction

Although currently there is a certain range of data available in various digital resources for researching the diversity of the Latvian language, for example, Latvian text corpora (www.korpuss.lv), they are often not useful for terminology research. The corpus included on this site – Balanced Corpus of Contemporary Latvian Texts – contains various texts with metadata, but it comprises only 10 million words and the ratio of scientific texts included in the corpus accounts for only 10% [1]. Of these 10%, only 2.5% are devoted to the field of biological science, and the authors of the corpus point out that this corpus is mostly developed for grammar research [2]. For example, when searching for the terminological name of the plant genus *Zinnia* in the Latvian language *cīnija (zinnia)*, only three sources with the use of the name of this plant genus are found. In two sources from regional and specialized periodicals – in the newspaper "Latgales Laiks" published on August 12, 2016 and in the gardening magazine "Dārza Pasaule" published in 2012, as well as in a work of fiction – the book "Sniega laika piezīmes" ["Snow Time Notes"] by Inga Ābele, published in 2004. Terminology research options in this corpus are also limited by the inclusion of only modern sources in the corpus database – i. e. starting from 1991, therefore it is not possible to compare changes in terms over a longer period of time. The scientific or in other words Latin name of this plant is not found in this corpus.

When searching for the Latvian name of the same plant genus name – *cīnija* in the largest electronic dictionary site available in Latvian – Tēzaurs (www.tezaurs.lv) it can be concluded that the entry for *cīnija* [3] gives a definition and two species names of this genus, and three different sources are given below, from which this information is compiled from – one is from the volume of the "Latviešu literārās valodas vārdnīca" ["Dictionary of the Latvian Literary Language"], one is from the volume "Latvijas enciklopēdija" ["Encyclopaedia of Latvia"] by the publishing house of Valērijs Belokoņs and one of the volumes of the encyclopaedia "Latvijas daba" ["Nature in Latvia"] [3], however, the affiliation of examples to the source used in this entry cannot be accurately determined and therefore all these sources have to be considered separately by the researcher.

When searching for the Latvian plant name of the genus *Zinnia*, National Terminology Portal (https://termini.gov.lv/) offers an entry from the Terminology Commission of the Latvian Academy of Sciences, the 9th term collection "Agronomijas terminu vārdnīca" ["Agronomic Term Dictionary"] [4], which indicates the Latvian name of the plant genus *Zinni*a – *cīnija* that is approved by the Terminology Commission. The metadata to this resource are accurate, however, it is also the only resource included in this portal with the name of that genus of plants. In practice, for example, the offer of the company 'Kurzemes sēklas" shows that seeds of these plant species are sold under the Latvian name of *cinnija* [5]. Also, in other specialized literature the Latvian name of the plant species *Zinnia elegans* is mentioned as *gleznā cinnija* (common zinnia or elegant zinnia) [6, p. 50]. In this case, it would be valuable to find out the usage frequency data of one or another variant of the Latvian name, as well as their fixation in different types of resources. For example, in Pauls Galenieks's work "Botāniskā vārdnīca" ["Botanical dictionary"] for the plant genus *Zinnia*, the given Latvian name is *cinijas* [7, p. 15]. These aggregated data already show that this Latvian name has changed over time. Such terminological studies require a specific selection of data. For example, research on the use of botanical or zoological terms, their changes or statistical research requires the use of diverse publications: scientific articles, periodicals, popular science articles, teaching and specialized literature, reference literature, laws and regulations, but these data are dispersed in different resources. Whereas, diachronic and etymology research requires further use of other digital resources, such as the ancient text corpus [8]. However, it only summarizes the texts of the 16th–18th century, therefore printed resources, as well as other additional digital resources, should also be used, including the data set in articles from the 19th and 20th centuries. In such studies, it would be important to use as diverse research material as possible in order to obtain a more accurate result of the study.

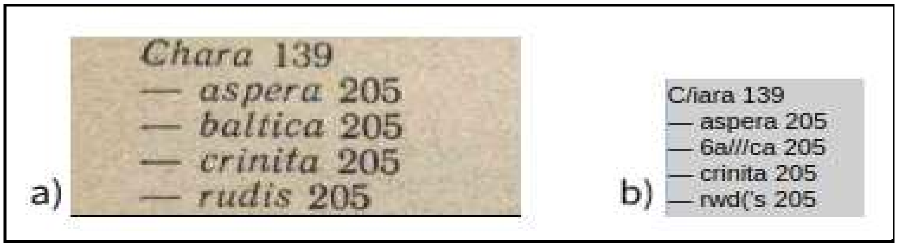
Only one example of an ornamental plant's name is described above. However, if one were to search for the terminological name of a medicinal plant, where human life might depend on the result, problems could occur. The principle of unambiguity in the terminology of Latvian plant names has been justified by Inese Ēdelmane [9]. This part of the lexicon is very important and it is crucial to select, structure and make it accessible, that would add significantly to the range of digital library resources. Other studies are also available on the wide scope of this special part of the lexicon, which makes the work of, for instance, lexicographers, more difficult [10].

There is no unified database for research of Latvian names for organisms, but Ventspils University of Applied Sciences, in cooperation with the Institute of Horticulture, is currently developing a special database for the research of the names of organisms (hereinafter IMDS) for this purpose. The new database will be with specific emphasis on the Latvian language and for research purposes, providing both historical information of the usage of lexemes, as well as terminological data source for a wide range of interested parties – translators, linguists and lexicographers, terminologists, authors. The terminological part of the database will be most important, and it will contain relations among scientific names and equivalent Latvian names. The structure, data retrieval and results of the database will be reported in future studies by the authors. Since the project will terminate at the end of 2023, the current scope of research is related to the automation of data entry. Most of the data (scientific names of organisms and their equivalents in other languages) are entered manually, but solutions for automated data addition are also considered. One of the solutions is the present study that proposes the processing of material with a structure of the dictionary and adding the data extraction to the database. The data retrieved by using the tool developed in this study shall be stored in this database, thus enabling the creation of a specific set of data for research purposes faster. The following chapters of this article provide an analysis of the developed tool set and a description of the methodology.

2. Research methodology

2.1. Characteristics of the Problem and the Tasks of the Study

The National Library of Latvia (hereinafter LNB) provides an opportunity access books not only to in printed version, but also in electronic form using the Latvian National Digital Library (hereinafter LNDL) tool (http://gramatas.lndb.lv/). The use of the above mentioned tool allows to create an IT solution that automatically processes LNB resources to supplement the database of organism names that is currently developed. The LNDL tool uses the Optical Character Recognition (hereinafter OCR) [11] algorithm to digitize scanned books, but often the used OCR algorithm has failed to convert the scanned image into a 100% correct text (see Fig. 1). Figure 1 shows fragments of the book "Augstāko augu sistemātika" ["Systematics of Higher Plants"] [12], which will be used further on to describe the methodology.



Chara 139
— aspera 205
— baltica 205
— crinita 205
— rudis 205

C/iara 139
— aspera 205
— 6a///ca 205
— crinita 205
— rwd('s 205

Figure 1: (a) Fragment of a scanned page of the book "Systematics of Higher Plants"; (b) result retrieved by LNDL tool.

Considering the inaccuracy of the results returned by the LNDL tool, own set of algorithms was developed for the study in order to automatically identify, process and store the names of organisms in the created database of organism names (IMDS) mentioned in the books provided by the LNDL tool. Following tasks were carried out in order to achieve the defined objective:

- Automated data retrieval i. e. retrieval of scanned book pages from the LNDL (http://gramatas.lndb.lv/);
- Preparation of scanned images of books for processing by an OCR algorithm;
- Processing images with an OCR algorithm that allows to retrieve information in text form;
- Evaluation and comparison of the results obtained and the results offered by the LNDL tool;
- Identification of the Latvian and Latin organism names in the obtained text.

Since the terminology of the field is not explicitly marked in the specialized literature, as is the case, for example, in the general translation dictionaries, where the terms are accompanied by the abbreviation of the industry designation: bot. – botany, zool. – zoology, etc., the developed solution should identify the scientific or Latin names of organisms that are often added to the respective organism names in texts from the field of biology. It is important to retrieve organism names in local languages (Latvian, Russian, German, etc.) by identifying them together with the given Latin names.

2.2. Used Technologies and Workflow of the Developed Solution

The developed methodology consists of five stages (see Fig 2):

(1) Book pre-processing stage, where the book is scanned as a pdf file and automatically converted to png files where each page is a unique png file. To retrieve LNDL book page representation in image format, Selenium architecture was used through Python library.

(2) scanned page processing stage, where multiple image processing methods are applied to the png images, for example, the background of the image is filtered using Otsu thresholding, then the small details of the image are retrieved using erosion and expansion operations, later the image quality is improved using erosion, reconstruction and closing morphological operations. For this stage, a pre-processing tool was developed in Python including Skimage, OpenCV2 and numpy libraries. Google Tesseract OCR technology which is based on LSTM (Long short-term memory) artificial recurrent neural network is used to process obtained png files and to digitize information from scanned png files into text document files where processed results are organized in multiple columns related to specific language. In this research, Tesseract OCR versions 3 and 4 were used. To combine all operations under one language, the PyTesseract and tesserocr libraries were used to enable the use of Tesseract using the Python environment.

(3) Language recognition stage, where obtained results of all names from the stage 2 are classified by Language identificator LangId. In this research only three data sets of languages are used – Latvian (LV), Latin (LA) and Russian (RU). Developed script in Python manages the language identificator process for each organism name in the text document, identifies unclear results and stores all obtained

results in the excel data file, where information about organism names are stored with information about language of organism name and commemoration page number.

(4) Result post-processing stage, where the user manually post-processes all results in the excel file. In this stage a developed tool in Python is used to integrate obtained results in the IMDS database.

(5) Data storage and displaying stage (IMDS), where users can view and process the integrated data. IMDS system consists of three main parts – (1) the database where MySQL technology is used, (2) the backend where Spring Framework technologies are used, and (3) the frontend where React JS technology is used.

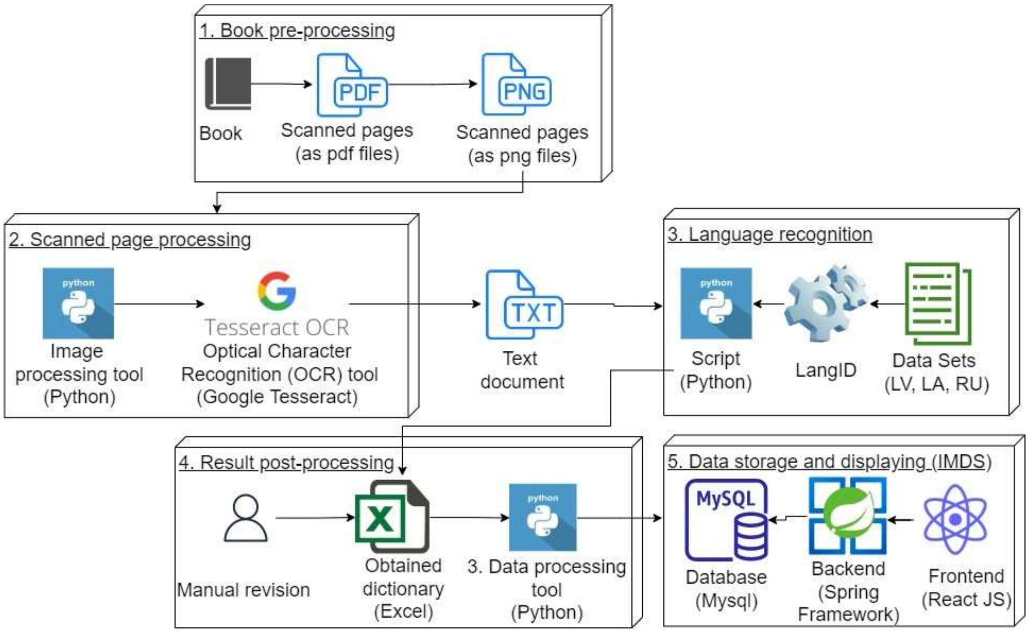

1. Book pre-processing
Book
Scanned pages (as pdf files)
Scanned pages (as png files)
PDF
PNG
2. Scanned page processing
Image processing tool (Python)
Tesseract OCR
Optical Character Recognition (OCR) tool (Google Tesseract)
TXT
Text document
3. Language recognition
Script (Python)
LangID
Data Sets (LV, LA, RU)
4. Result post-processing
Manual revision
Obtained dictionary (Excel)
3. Data processing tool (Python)
5. Data storage and displaying (IMDS)
MySQL
Database (Mysql)
Backend (Spring Framework)
Frontend (React JS)

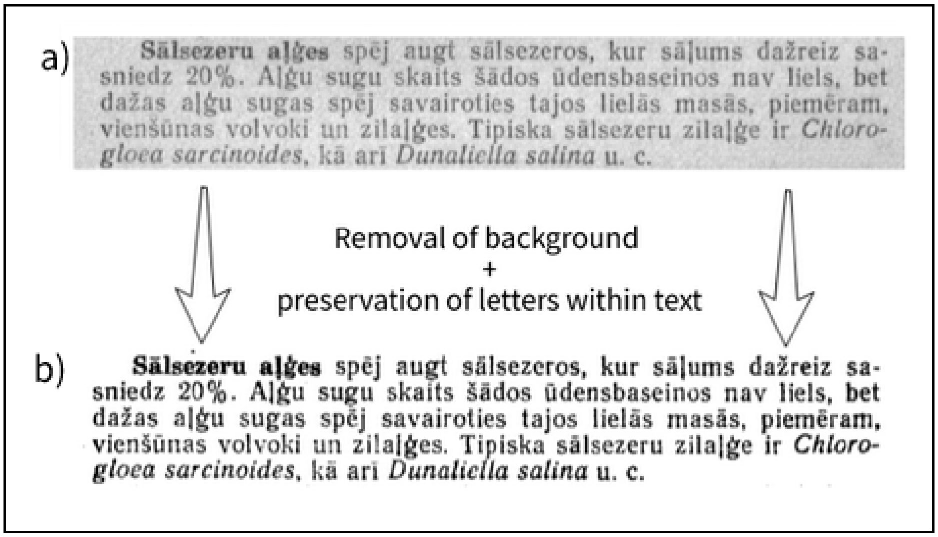
Figure 2. Used technologies and workflow of the developed solution.

2.3. Processing and Digitization of Scanned Images

For more accurate results, primary processing of the images had to be carried out before performing the OCR operation. Depending on the problem, the actions described further on might vary, but the objective outlined in this study is to obtain a binary image with as little light gradient effect as possible. Ideally, the pixels belonging to the text in the image array would be marked with true values, and the background – with false values.

Given that scanned books do not always have a homogeneous background and there are various noises from lighting and material, it is necessary to use additional processing algorithms to create a binary image. A popular solution to these problems is the use of morphological methods of images [13] – moving the structural element through the image matrix, performing dilation or erosion operations at each step. These basic operations can be combined to create new ones, such as opening and closing operations. The methodology provides an opportunity to improve the quality of letters, retrieve the small details of the images, as well as to improve the possibility of higher quality image thresholding.

Thresholding in the context of image processing means evaluating pixels to determine if they meet criteria specified by the user. Since colour information is not as important for this task, it is more appropriate to use the intensities of grey tone values. In simple cases, it is possible to specify a threshold chosen by the user, but given that the scanned pages have different lighting, it is necessary to automatically calculate the threshold values. The algorithm uses the Otsu [14] thresholding method, which calculates the threshold based on keeping the sum of foreground and background intensity dispersion to a minimum.


a)

Sālsezeru aļģes spēj augt sālsezeros, kur sāļums dažreiz sasniedz 20%. Aļģu sugu skaits šādos ūdensbaseinos nav liels, bet dažas aļģu sugas spēj savairoties tajos lielās masās, piemēram, vienšūnas volvoki un zilaļģes. Tipiska sālsezeru zilaļģe ir *Chlorogloea sarcinoides*, kā arī *Dunaliella salina* u. c.

Removal of background
+
preservation of letters within text
b)

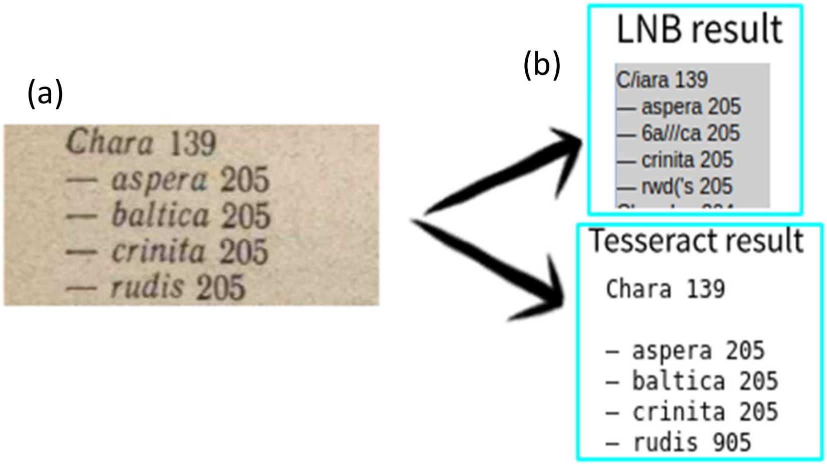
Sālsezeru aļģes spēj augt sālsezeros, kur sāļums dažreiz sasniedz 20%. Aļģu sugu skaits šādos ūdensbaseinos nav liels, bet dažas aļģu sugas spēj savairoties tajos lielās masās, piemēram, vienšūnas volvoki un zilaļģes. Tipiska sālsezeru zilaļģe ir *Chlorogloea sarcinoides*, kā arī *Dunaliella salina* u. c.

Figure 3 (a) Example of a scanned page; (b) result obtained by image processing algorithm.

Examples in the Figure 3 and the first results obtained in the study lead to the conclusion that the result obtained by the algorithm is clearer and that the text contains clearly visible words and language units to be retrieved (see Fig. 3).

For the OCR algorithm, the Google Tesseract technology was used because of its promising results in the similar use cases related to text digitization from books [15] [16][17]. Obtained results showed that Tesseract technology usage was good enough to identify the names of the organisms also in this case, so no other OCR tools were used for further research, however Keras-OCR, Amazon Textract, Google Document AI and EasyOCR should be mentioned as options.

After the image processing stage, the obtained result is a binary image, which is then automatically processed with the Tesseract OCR tool to retrieve textual information that can be automatically processed retrieving the organism names mentioned in the text. Figure 4 below shows the result obtained by the LNDL tool and the result of the algorithm developed in this study.



Chara 139
— *aspera* 205
— *baltica* 205
— *crinita* 205
— *rudis* 205

LNB result

```
C/iara 139
— aspera 205
— 6a///ca 205
— crinita 205
— rwd('s 205
```

Tesseract result

```
Chara 139

- aspera 205
- baltica 205
- crinita 205
- rudis 905
```

Figure 4: (a) Example fragment of a scanned page; (b) result obtained by image processing and OCR algorithm: above – LNDL tool, below – algorithm developed in this study.

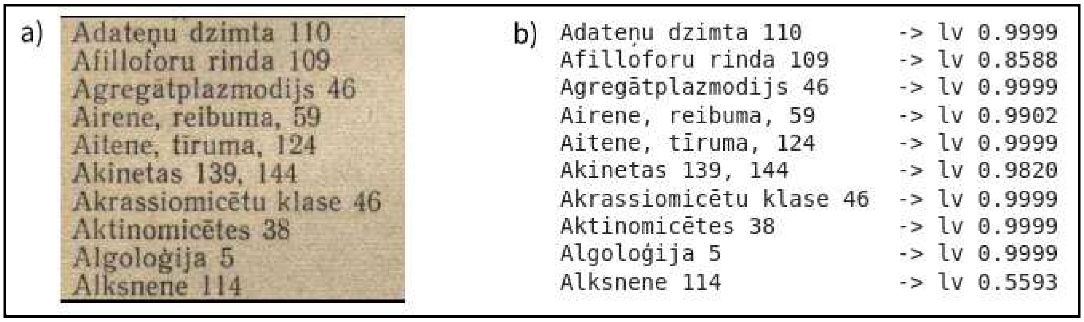
Figure 4 shows that the developed algorithm makes a good distinction between graphemes, whereas the results of the LNDL tool are converted, not relevant and cannot be identified.

2.4. Recognition of Language for Retrieved Words

Next step is to process this obtained research material with a language identification algorithm. For this step multiple tools for language detection/recognition were used, for example, (1) Natural Language Toolkit with Language ID module using TextCat algorithm; (2) pycld3 package which contains Python bindings (via Cython) to Google's CLD3 library; (3) Polyglot which depends on pycld2 library and on cld2 library for detecting language(s) used in plain text; (4) TextBlob that provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis; (5) LangId [18] that contains a trained neural network and a 97-language trainee dataset. LangId showed the best result for Latvian and Latin languages recognition and it was used for future applications. In this research, methods that do not require additional training were chosen, making them more versatile / use-case independent.

The LangId tool can be retrained according to the needed precision in the execution of a specific task. To identify the conformity of the specific organism names in Latvian, Russian or Latin included in the text fragment, multiple data sets were used to examine the methodology – the text in Latvian, the text in Latin, the text in Russian, the text containing Latvian and Latin words, the text containing all three languages – Latvian, Latin and Russian. The identification algorithm itself is based on the neural network which is trained using corpus data. Various field specific texts have been used in Latvian, Russian and Latin to train the algorithm.

The methodology is represented in the examples in Figures 5, 6, 7 and 8, showing instances of the text fragments, thus reflecting the effectiveness of the methodology and the processing results obtained for specific test data. The results provide a probability of matching each word in Latvian, Russian and Latin in each of the data sets. Probability is shown in the 0–1 scale where 1 corresponds to 100% confidence matching.



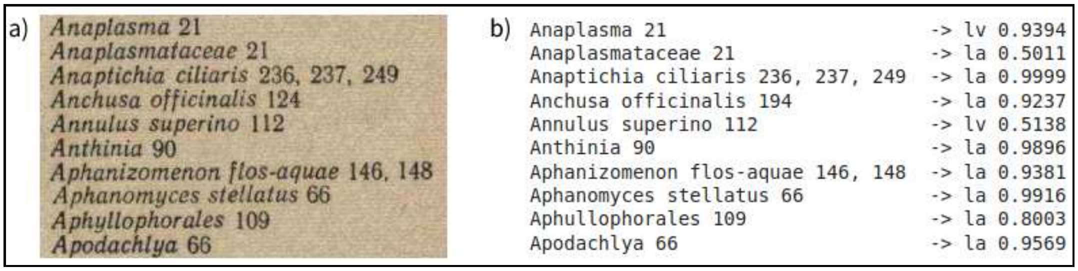
a)

Adateņu dzimta 110
Afilloforu rinda 109
Agregātplazmodijs 46
Airene, reibuma, 59
Aitene, tīruma, 124
Akinetas 139, 144
Akrassiomicētu klase 46
Aktinomicētes 38
Algoloģija 5
Alksnene 114

b)

```
Adateņu dzimta 110         -> lv 0.9999
Afilloforu rinda 109       -> lv 0.8588
Agregātplazmodijs 46       -> lv 0.9999
Airene, reibuma, 59        -> lv 0.9902
Aitene, tīruma, 124        -> lv 0.9999
Akinetas 139, 144          -> lv 0.9820
Akrassiomicētu klase 46    -> lv 0.9999
Aktinomicētes 38           -> lv 0.9999
Algoloģija 5               -> lv 0.9999
Alksnene 114               -> lv 0.5593
```

Figure 5: Application of the methodology developed in the study data set with text in Latvian: (a) Fragment of a scanned page; (b) Probability of compliance of each digitized language unit in Latvian and Latin.



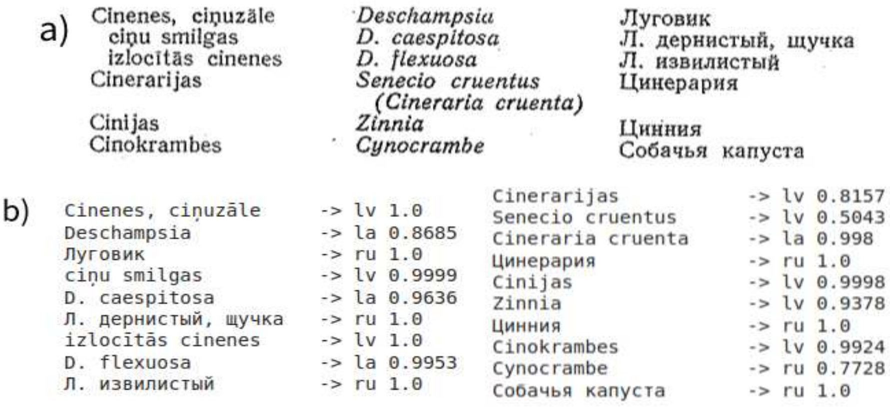
a)

Anaplasma 21
Anaplasmataceae 21
Anaptichia ciliaris 236, 237, 249
Anchusa officinalis 124
Annulus superino 112
Anthinia 90
Aphanizomenon flos-aquae 146, 148
Aphanomyces stellatus 66
Aphyllophorales 109
Apodachlya 66

b)

```
Anaplasma 21                              -> lv 0.9394
Anaplasmataceae 21                        -> la 0.5011
Anaptichia ciliaris 236, 237, 249         -> la 0.9999
Anchusa officinalis 194                   -> la 0.9237
Annulus superino 112                      -> lv 0.5138
Anthinia 90                               -> la 0.9896
Aphanizomenon flos-aquae 146, 148         -> la 0.9381
Aphanomyces stellatus 66                  -> la 0.9916
Aphullophorales 109                       -> la 0.8003
Apodachlya 66                             -> la 0.9569
```

Figure 6: Application of the methodology developed in the study to a data set containing only Latin text: (a) Fragment of scanned page; (b) Probability of conformity of each digitized language unit with Latvian and Latin.



a)

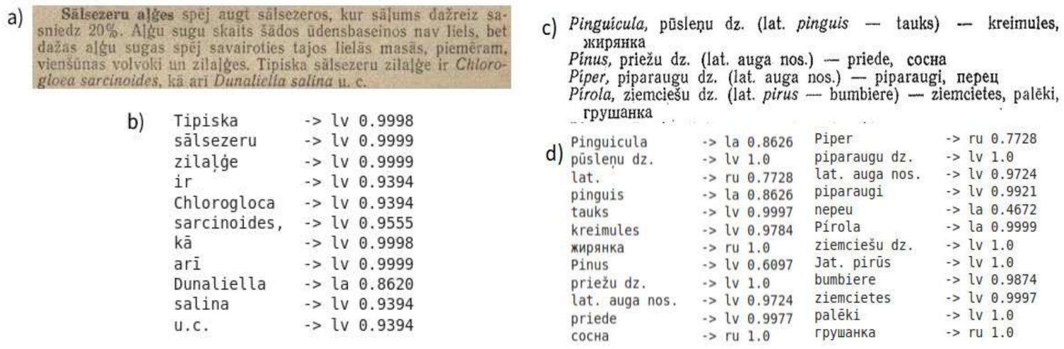
Cinenes, ciņuzāle	*Deschampsia*	Луговик
ciņu smilgas	*D. caespitosa*	Л. дернистый, щучка
izlocītās cinenes	*D. flexuosa*	Л. извилистый
Cinerarijas	*Senecio cruentus* (*Cineraria cruenta*)	Цинерария
Cinijas	*Zinnia*	Цинния
Cinokrambes	*Cynocrambe*	Собачья капуста

b)

```
Cinenes, ciņuzāle       -> lv 1.0
Deschampsia             -> la 0.8685
Луговик                 -> ru 1.0
ciņu smilgas            -> lv 0.9999
D. caespitosa           -> la 0.9636
Л. дернистый, щучка     -> ru 1.0
izlocītās cinenes       -> lv 1.0
D. flexuosa             -> la 0.9953
Л. извилистый           -> ru 1.0
```

```
Cinerarijas             -> lv 0.8157
Senecio cruentus        -> lv 0.5043
Cineraria cruenta       -> la 0.998
Цинерария               -> ru 1.0
Cinijas                 -> lv 0.9998
Zinnia                  -> lv 0.9378
Цинния                  -> ru 1.0
Cinokrambes             -> lv 0.9924
Cynocrambe              -> ru 0.7728
Собачья капуста         -> ru 1.0
```

Figure 7: Application of the methodology developed in the study to a data set from P. Galenieks's work "Botanical Dictionary": (a) Fragment of scanned page where plant names organized in three columns – Latvian, Latin and Russian; (b) Probability of conformity of each digitized language unit with Latvian, Latin and Russian.



a)

Sālsezeru aļģes spēj augt sālsezeros, kur sāļums dažreiz sasniedz 20%. Aļģu sugu skaits šādos ūdensbaseinos nav liels, bet dažas aļģu sugas spēj savairoties tajos lielās masās, piemēram, vienšūnas volvoki un zilaļģes. Tipiska sālsezeru zilaļģe ir *Chlorogloea sarcinoides*, kā arī *Dunaliella salina* u. c.

b)

```
Tipiska         -> lv 0.9998
sālsezeru       -> lv 0.9999
zilaļģe         -> lv 0.9999
ir              -> lv 0.9394
Chlorogloca     -> lv 0.9394
sarcinoides,    -> lv 0.9555
kā              -> lv 0.9998
arī             -> lv 0.9999
Dunaliella      -> la 0.8620
salina          -> lv 0.9394
u.c.            -> lv 0.9394
```

c)

Pinguícula, pūsleņu dz. (lat. *pinguis* — tauks) — kreimules, жирянка
Pínus, priežu dz. (lat. auga nos.) — priede, сосна
Píper, piparaugu dz. (lat. auga nos.) — piparaugi, перец
Pírola, ziemciešu dz. (lat. *pirus* — bumbiere) — ziemcietes, palēki, грушанка

d)

```
Pinguicula        -> la 0.8626    Piper             -> ru 0.7728
pūsleņu dz.       -> lv 1.0       piparaugu dz.     -> lv 1.0
lat.              -> ru 0.7728    lat. auga nos.    -> lv 0.9724
pinguis           -> la 0.8626    piparaugi         -> lv 0.9921
tauks             -> lv 0.9997    nepeu             -> la 0.4672
kreimules         -> lv 0.9784    Pírola            -> la 0.9999
жирянка           -> ru 1.0       ziemciešu dz.     -> lv 1.0
Pinus             -> lv 0.6097    Jat. pirūs        -> lv 1.0
priežu dz.        -> lv 1.0       bumbiere          -> lv 0.9874
lat. auga nos.    -> lv 0.9724    ziemcietes        -> lv 0.9997
priede            -> lv 0.9977    palēki            -> lv 1.0
сосна             -> ru 1.0       грушанка          -> ru 1.0
```

Figure 8: Application of the methodology developed in the study to a data set containing text in multiple languages: (a) Fragment of a scanned page where data set containing text in Latvian and Latin; (b) Probability of conformity of each digitized language unit with Latvian and Latin; (c) Fragment of a scanned page where data set containing text in Latvian, Russian and Latin; d) Probability of conformity of each digitized language unit with Latvian, Russian and Latin.

As shown in Figure 5, when working with a data set containing words only in Latvian, the algorithm provides a 100% correct match in Latvian, thus it would be possible to process Latvian language dictionaries and automatically save the retrieved language units in a developed database. Results obtained in the example displayed in Figure 6 show that only 80% of Latin words are correctly identified, so it would not be possible to save this data to the database automatically, since verification and manual editing of the results is required. For all the data that has a probability of compliance below a certain threshold (e. g. 80%) a manual check would be applied as a post-editing technique. In the example in Figure 6, the first word *anaplasma* (anaplasma) is identified as a Latvian word with a probability of 93%, although *anaplasma* is a Latin origin word. It is possible that the algorithm has linked it to the Latvian *plazma* (plasma), which differs from the Latin word only by one different letter – *z*. When testing the algorithm with other text fragments in Latin, the accuracy in several cases was even lower than 80%. Another reason for the algorithm's inaccuracy could be that a Latin data set containing few texts on biology and botany was used to train the neural network.

Working with a data set that includes Latvian, Russian and Latin languages, plant names in Russian were identified with approximately 100% precision as it is shown in Figure 7. This is how the *Zinnia* example described in the introduction of the article is also represented in the book. However, the Latin name has been incorrectly determined by the algorithm, even though there is a 4 character difference between the Latin and Latvian names. Despite the algorithm identifying plant names in Latvian with high precision, it also identified some of the plant names in Latin as plant names in Latvian, and it can be trained using biology data sets in Latvian and Latin, however this specific set of biology data may not be sufficient enough for neural network training. [19] The developed algorithm works correctly with plant names in Russian and Latvian, but since Latin and Latvian use the Latin alphabet, the algorithm needs improvements for the distinction of both languages.

Figure 8 (a) and (b) shows an example of text that contains both Latvian and Latin words, and the results are similar to those of the previous two cases – the algorithm has identified Latvian words with 100% accuracy, while the Latin words have a 50% match accuracy. Obtained result is also similar when processing other equivalent text fragments – the word match accuracy for Latvian is close to 100%, but the word matching accuracy for Latin is 30% to 50%. After processing the third data set, which contains words in both Latvian and Latin, it must be concluded that it is necessary to retrain the neural network with a larger set of training data in Latin, which would include more texts from scientific literature in the field of biology.

Figure 8 (c) and (b) shows an example of the text that contains Latvian, Russian and Latin words. If the OCR algorithm recognized the scanned image into its correct digitized version, words in Russian were identified correctly. Language identification algorithm working in the Russian language is promising because the probability of errors is low in all cases. It can be explained by the Russian alphabet being unique among all used languages in this study and it is easy to differentiate because of the Cyrillic symbols. However, words in Latin and Latvian could be confused. Currently the language identification algorithm works best when the alphabets are different.

3. Research Results

The methodology developed in the study and described above was tested using the book "Systematics of Higher Plants" by V. Langenfelds, E. Ozoliņa and G. Ābele, and the test results show that the methodology can successfully digitize poor quality scans and identify the conformity of the text in Latvian, but there are problems with the identification of Latin text. Given the specific features of this methodology, the application of the methodology could work better when digitizing dictionaries, because words are arranged in columns there and matching could be done for the respective languages. The dataset in the books selected before was too small to draw convincing conclusions for testing the methodology, so another book was analysed for the study.

Data processing from the P. Galenieks's work "Botanical Dictionary" was carried out according to the methodology described before. It should be noted that this dictionary is not available in LNB's catalogue of digitized books, nor is its data added, for example, in the Latvian National Terminology portal (https://termini.gov.lv/). First, each page of the book was scanned, then data from chapter 1 of the dictionary, p. 5–70, was retrieved and analysed. Figure 9 in the article shows the level of certainty about the correctness of the plant names language recognition in each page. It is possible to retrieve precision data from the algorithm for each analysed element, however, it may be inaccurate, as the analysed element may not be a word, but a blank space or so called "noises" (e. g., dots, lines, needless symbols, etc.). Therefore, the image (see Figure 9) highlights the average certainty of the page. This figure also shows linear regression curves for each graphic (dotted red, purple and orange lines).

When processing data from Chapter 1 of the book with a language identifier, data on 2204 plant names in Latvian, 2135 in Russian, 2287 in Latin were obtained. Russian plant names gave 100% correctness. It is not surprising, because Russian alphabet differs the Latin and Latvian alphabets and it is easy to identify Cyrillic symbols. When the processing algorithm had to choose between Latvian and Latin, identification success of Latvian was 90.6%, however only 39.74% of Latin words were identified correctly.

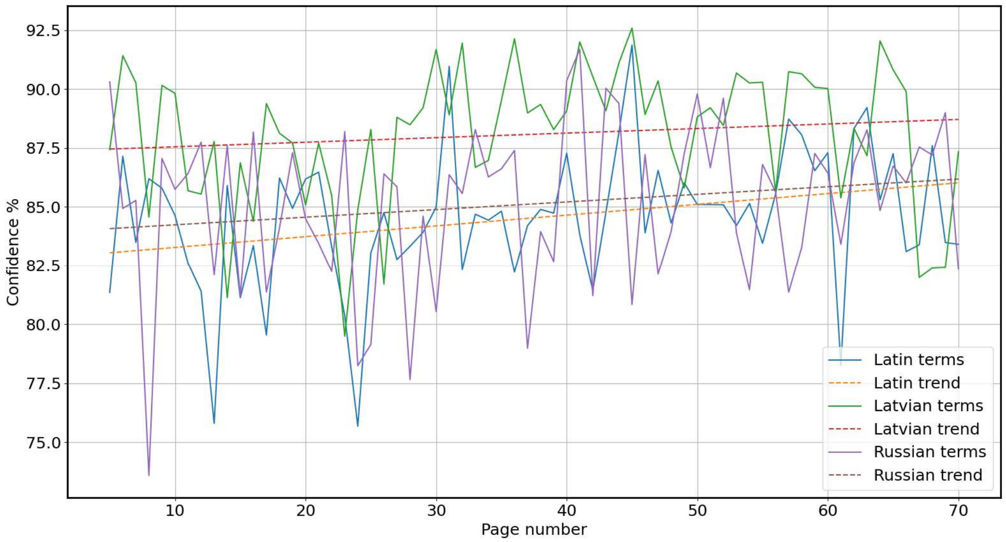


Figure 9: Results of data reading with the algorithm developed in the study from P. Galenieks's "Botanical Dictionary", p. 5–70.

After choosing a certain section from P. Galenieks's "Botanical Dictionary", a total of 65 pages of text were processed with the developed solution. That is a main part of the dictionary, and plant names are arranged in three columns like in dictionaries. In the processed 65 pages, overall 6626 plant names were organized in three columns – Latvian, Latin and Russian. This format was chosen for checking the accuracy of the language identification algorithm. The results of recognizing the text from the images are so good that they can be used for retrieval of the data and addition to the database. Achieving full precision of the data retrieved from the study requires human involvement, as the data should be carefully reviewed, compared with the printed work and corrected if necessary. Manual processing of data from Galenieks's dictionary (text review and collation, correcting errors, adding page numbers) after extraction requires 0.5 staff-hour for 10 pages or 0.0024 man-hour per entry consisting of one scientific plant name and 2 local language names, in total 0.0008 staff-hour per one plant name. To compare our developed software solution with data entry processes done by persons, manual entry actions of six people were analysed during September–December 2021. During this period six people stored 20704 organism names in the developed IMDS database. Taking into account the staff-hours spent to enter organisms names, obtained results showed overall 0.0263 staff-hour per one organism name entered manually, which is 32 times slower than using our developed solution.

4. Conclusions

The study described in the article is part of a larger study developing a database of organism names (IMDS) whose main task is to address the dispersion of data across different digitized and non-digitized resources, as well as to produce data with accurate metadata that would be useful for research purposes.

These data processing operations have been performed using existing tools in order to develop a simplified processing workflow for this type of dictionaries. The methodology developed in this study could be useful for digitizing other books (including those with a dictionary structure), where the local language names of plants, animals (also terms from other fields, e. g. medical terms in local language and Latin) are linked to their scientific or Latin name. The methodology developed in the study should be a new contribution to the further digitization and processing of specialized literature for the study of texts in this field as well.

It can be concluded that the Latvian language and Russian language models are sufficiently well trained to correctly identify the conformity of language units with the Latvian and Russian languages. The existing Latin model, on the other hand, is not sufficiently well trained to identify terms in the field of botany. The Latin names of the taxa of organisms often are given in italics in the printed books. This is also the case in P. Galenieks's work "Botanical Dictionary", where all Latin names are in italics making data reading from images more difficult. This means worse quality input data for language identification. Also, similar alphabets in Latvian and Latin could make it more difficult to identify words in Latin.

In perspective, the approach described in this study could be applied to other work in similar fields, such as the digitization of dictionaries of zoological terms. Similarly, medicine is one of the fields where Latin terms are often used in addition to local terms.

The data prepared by the method described in the study will be further collected in the IMDS database. However, the data acquisition methodology described and developed in the study also relieves the manual work of the data entry personnel when entering data in the intended database. The overall process takes less time than entering all organism names manually. Obtained results of calculations showed that using developed software and manually checking is 32 times faster than manual entry in the IMDS database.

5. Acknowledgements

This research has been funded by the Latvian Council of Science, project "Smart complex of information systems of specialized biology lexis for the research and preservation of linguistic diversity", No. lzp-2020/1-0179.

6. References

[1] Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss (LVK2018 (beta). [The Balanced Corpus of Modern Latvian.] University of Latvia – Institute of Mathematics and Computer Science, 2018. URL: http://www.korpuss.lv.

[2] K. Levāne-Petrova, Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos [The Balanced Corpus of Modern Latvian, its role in grammar studies.], in A. Kalnača (Ed.), Valoda: nozīme un forma, 10. Latvijas gramatiskā doma gadsimta gaitā [Language: Meaning and Form, 10: A century of Latvian grammatical thought.], Riga: Latvijas Universitātes Akadēmiskais apgāds, 2019, pp. 131–146.

[3] A. Spektors, The dictionary and thesaurus of Latvian Tēzaurs.lv. URL: https://tezaurs.lv/cīnija.

[4] Latvijas Nacionālais terminoloģijas portāls Termini.gov.lv. [National Terminology Portal Termini.gov.lv.] URL: https://termini.gov.lv/atrast/cīnija.

[5] Kurzemes sēklas, SIA. Sēklu piedāvājuma elektroniskais katalogs. [Electronic catalogue of seed supply]. URL: https://www.kurzemesseklas.lv/lv/product/759-cinnija-aztec-sunset.

[6] I. Birulis, 400 puķes Latvijā no pavasara līdz rudenim. [400 flowers in Latvia from spring to autumn], Riga: AS "Lauku Avīze", 2007.

[7] P. Galenieks, Botāniskā vārdnīca. Rīga: Latvijas Valsts izdevniecība, 1950.

[8] Latviešu valodas seno tekstu korpuss SENIE. [The Corpus of Ancient Latvian texts SENIE], URL: http://www.korpuss.lv/id/Senie

[9] I. Ēdelmane, Ārstniecības augu terminoloģija, in Latviešu valodas kultūras jautājumi. Volume 4. Rīga: Liesma, 1968, pp. 103–107.

[10] F. Dornseiff, Der deutsche Wortschatz nach Sachgruppen. 8. völlig neu bearb. Aufl. von Uwe Quasthoff, Berlin, New York: de Gruyter, 2004.

[11] L. Eikvil, Optical Character Recognition, Norsk Regnesentral, Oslo, Norway, Rep. 876, 1993.

[12] V. Langenfelds, E. Ozoliņa, G. Ābele, Augstāko augu sistemātika. [Systematics of Higher Plants], Rīga: izdevniecība "Zvaigzne", 1973.

[13] N. Efford, Digital Image Processing: A Practical Introduction Using JavaTM. Pearson Education, 2000.

[14] S. L. Bangare, A. Dubal, P. S. Bangare, S. T. Patil, Reviewing Otsu's Method For Image Thresholding, International Journal of Applied Engineering Research 10(9), 2015/1, pp. 21777–21783.

[15] T. Hegghammer, OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment, Journal of Computational Social Science, 2021.

[16] C. Clausner, A. Antonacopoulos, S. Pletschacher, Efficient and effective OCR engine training, International Journal on Document Analysis and Recognition (IJDAR), Volume 23, 2020, pp. 73–88.

[17] J. Martínek, L. Lenc, P. Král, Building an efficient OCR system for historical documents with little training data, Neural Computing and Applications. Neural Computing and Applications, Volume 32, 2020, pp. 17209–17227.

[18] M. Lui, T. Baldwin, langid.py: An Off-the-shelf Language Identification Tool, Proceedings of the ACL 2012 System Demonstrations, 2012, pp. 25–30.

[19] R. Darģis, K. Levāne-Petrova, I. Poikāns, Lessons Learned from Creating a Balanced Corpus from Online Data, Human Language Technologies – The Baltic Perspective, Volume 328, 2020, pp. 127–134.