Automatic Detection of Dates in the Corpus of Diaries

Haralds Matulis[1], Sanita Reinsone[1], and Ilze Ļaksa-Timinska[1]

[1] *Institute of Literature, Folklore and Art of the University of Latvia, Mūkusalas 3, Riga, LV1423, Latvia*

Abstract

This paper deals with the automatic detection of dates in a corpus of digitized, hand-written diaries in Latvian. Date detection is an important step in processing diaries' corpus, as it allows to split the source texts by dates of entries and carry out diachronic analysis for separate diaries and compare metrics across different authors. This paper describes the workflow of data processing, provides step by step implementation of date detection algorithm, and gives an evaluation of empirical results with discussions of encountered practical challenges for precise date detection in personal diaries.

Keywords

date detection, corpus analysis, crowdsourcing, digitization, hand-written texts.

1. Introduction: Pilot Corpus of Diaries

Diary writing tradition is a complex phenomenon [4]. Forms and styles of how personal diaries are written can differ even within one notebook of a single author. However, it can be assumed that the date at the beginning of a daily record is a formal element that distinguishes diaries from other forms of personal autobiographical reflections in written form. Dates are important formal elements that keep diaries structured and aligned with the narrated time.

The corpus of diaries was built by the Institute of Literature, Folklore and Art (ILFA), University of Latvia. It is based on the Autobiography Collection, started in 2018, in which various life writing materials are archived and made digitally accessible at http://autobiografijas.lv. In collecting materials, priority was given to previously unpublished autobiographical items and to those not housed in other cultural heritage institutions: in other words, to those generally inaccessible because they were stored at home. At the next stage, a crowdsourcing platform at garamantas.lv was used to transcribe diaries from the scanned manuscripts. The file transformation from hand-written to electronic text was performed (1) sometimes by the authors of the diaries themselves, (2) occasionally by the project researchers at ILFA, (3) and mostly by volunteers [6]. The collection covers the 20th century through today. For the pilot corpus, 36 diaries were selected which were fully transcribed. They are of different lengths (some diaries consisting of a few entries, others written over as many as 55 years), written by authors of different age and educational and social backgrounds in the Latvian language from 1917 to 2021.

Information on the creation of corpora of diaries and the use of digital methods in its analysis is relatively scarce in the academic literature, and studies on this topic are not many. Existing research articles have mostly focused on the creation of publicly accessible digital editions of diaries, for example, reflecting upon the encoding of diaries in a particular format or annotating a transcript with persons, places, and other named entities ([8] and [5]). A relatively recent study has developed a new system for the computer-aided identification of narrative threads in diary-like online blogs, using several natural language processing techniques [2]. A team of researchers at the University of Adelaide has been building a corpus of World War I diaries, containing over 500 diaries written between August 1914 and November 1918. Applying a number of distant reading methods, the study provides a general overview, showing thematic trends and the distribution of particular concepts across the corpus [1]. To date, we have not been able to find studies that have analyzed the creation of a comprehensive corpus

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022
EMAIL: haralds.matulis@gmail.com (A. 1); sanita.reinsone@lulfmi.lv (A. 2); ilze.laksa-timinska@lulfmi.lv (A. 3)
ORCID: 0000-0002-0142-7677 (A. 1); 0000-0003-1980-5450 (A. 2); 0000-0001-7213-4954 (A. 3)
© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3232/paper32.pdf

of diary texts or studies where such a corpus, potentially heterogeneous and yet representative, has been analyzed with computational methods.

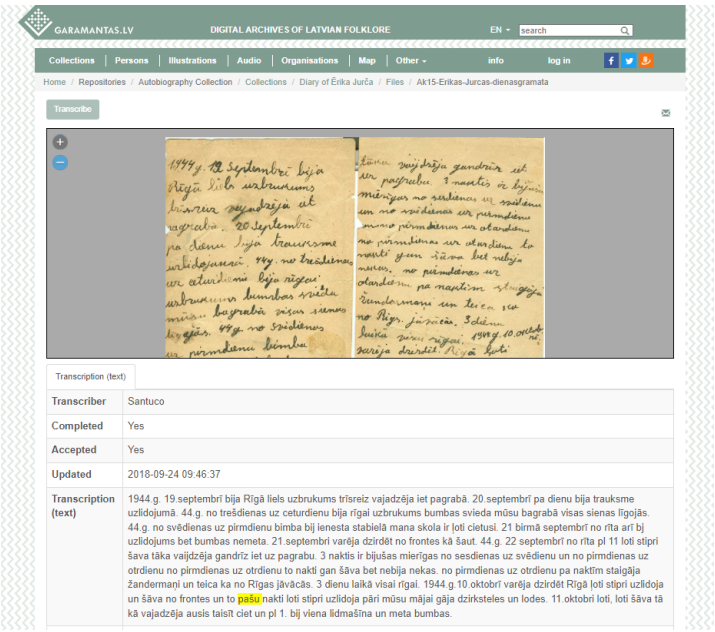

Figure 1: Public interface of the manuscript transcription tool at garamantas.lv.

An analysis of the date entries in the diary text corpus, with the aim of identifying the system of date notation in each individual diary, revealed that date notation tends to be a highly creative process, the author's taste, habits, and mood playing important roles. The date may be written at the beginning or at the end of the entry, sometimes in the middle; sometimes, too, the entry may be without a date because it is contextualized in the text. There are authors who mention the year only at the beginning of the year, and the month at the beginning of the month, numbering the days with numerals. The abbreviations of the year and the months are also very varied, and the use of punctuation and separators (full stop, comma, slash, colon, semicolon, hyphen) is also varied. Finally, there are also errors in months, days, or years. The landscape of dates in the diaries is indeed colorful.

2. Methodological Considerations on Date Detection in Diaries

For the further analysis of diaries with different computational methods — topic modeling, change of topics and sentiment over time, and comparison of metrics across different diaries — there was a need for a more detailed breakdown of the source files. The decision was to slice every diary into smaller chunks, extracting entries for single days. A single day is a semantically meaningful time unit for analysis, when compared with a larger time period like a week or a month or a finer unit such as morning, afternoon, or evening. Single day also coincides with the dominant notation system used by diary authors to register their writings. In this corpus there are only some exceptions when a diary's entries are undated or refer to a longer time periods, like a month or a season. Therefore, a day seems a reasonable choice for the finer partition of diaries.

As data after this pre-processing would be used in humanities research, the precision of the data is crucial; a decision was made to target initially the maximum data, possibly erring on the side of too many false positives. In the next step, the output file was given to a digital humanities researcher, who examined and manually corrected the wrong dates and deleted the non-dates so that the final version of the digitized diaries would be close to perfect accuracy. As the date-detection process was conducted in a cooperation between humanities researchers and a data scientist, the workflow had to be comprehensible and simple for both sides. The source files were text files (.txt format), which were shared on a drive. The data scientist downloaded the files, processed them, and uploaded the processed files back onto the drive in a separate folder. The input and output files were both .txt format, thus enabling both parties to access and work with files.

The specific challenge of finding dates in the diary corpus consisted of two parts. First, to find all dates occurring in the corpus. This was not an easy task, as the text was primarily meant for personal use and not for computational analysis. Therefore, the style of date notation is oftentimes elliptic, sometimes obscure, and varies widely even within one diary written by the same author. Second, to find all dates which are serving as *metadates* to denote the day, month, and year of the specific entry — and to distinguish these dates from false positives, i.e., such dates that only refer to some moment in the narrative but do not indicate the time when this record was made.

The wide variety of *metadate* notation, even by one author, might seem puzzling at first. And it might provoke a question: are authors deliberately negligent or obscure with date formats? The answer is that diaries, at least from the start, are written for oneself, and such elliptic ways of *metadate* formatting are sufficient for the author and his/her purposes.

31. 12. 32.	*3/I. 33.*	*2. jūlijā 1933. g.*	*7. septembra naktī*	*10. Septemb. 1933. g.*	*5. Septembrī*
š. g. 11. septm.	*1935. g.*	*1933. gadā 1. janvārī*	*2. janvārī 1935. g.*	*31. Oktobrī 1933. gadā*	*31/V. 36. g.*
12. aug. 1938. g	*6. jūlijā 1941. g.*	*1. 3. 44. g.*			

Figure 2: Many types of date notation in the diary of Davis Dauvarts, LFK Ak145.

Another observation is that a change in *metadate* format usually occurs with a larger time interval between records — that could be as long as several years or as short as a month, after which the author chooses to record the *metadate* in a format which seems more natural at that moment (perhaps not reviewing the diary to compare the previous format used). Maybe such variety in one diary indicates that writing is not the main or a prominent part of this person's day and lifestyle, as daily writing habits tend to develop more uniform patterns in a person's writing.

3. Different Conventions of Date Notation and Placement

In regard to their *metadate* notation system, all diaries of this corpus can be divided into two groups: those with an *absolute* and those with a *relative* system of *metadate* notation. The vast majority of diaries fall into the group of *absolute* date. By *absolute* we refer to a notation system where the date, in a full or shortened version, appears in a diary: dd.mm.yyyy – 14.02.1957; dd.mm.yy – 14.02.57; dd.mm – 14.02. A *relative metadate*, on the other hand, relies on the overall hierarchical structure of the *metadate* notation system in the diary, and the precise date of the entry can be deduced from the position of that entry in the overall structure of the diary. Of all diaries, there were only two diaries using a *relative date* system; these were addressed separately, by devising a particular search algorithm for each one. All further discussion about date detection is about *absolute metadate* detection.

The most frequent placement of the *metadate* adheres to this convention containing three rules: (1) an empty line before an entry of a new day, (2) the *metadate* is written at the beginning of a new line, (3) the diary entry for that day starts on the next line. However, there were variations to this system, which had to be accounted for.

After doing some pilot experiments for *metadate* detection and evaluating the results, it soon became clear that date_and_month is the most important part of the full_date record, as both the date and the month are needed to determine the precise entry time of the diary record. Without 'month' we are left just with the day — wondering in which month the author made that record. Without 'day' we are left with just the month, unable to attribute this entry to a specific dd.mm.yyyy. However, without a 'year,' we can still usually guess what year it is from the context and previous records.

As date_and_month is the critical minimum of information needed to find and recognize if that line contains the *metadate*, all input files were searched for this group. Although date_and_month group of the *metadate* is usually placed at the beginning of the entry and on a new line, it is not always at the *exact* beginning of the line.

First, there could be a year before date_and_month: 1957. gada 14. februārī.

There could also be a weekday before: Ceturtdien, 14. februārī.

Or a location of the entry: Stokholmā, ceturtdienā, 14. februārī.

In some less common cases the date_and_month is intertwined with the words of the first sentence: Ir jau pienācis 14. februāris… // And so the 14th of February has already come …

Therefore, a buffer of 25 characters was allocated to the beginning of the paragraph, allowing for the date_and_month group to start anywhere from the 1st through 25th character of the paragraph, but not later. Practical experiments with larger intervals showed not to improve date detection quality while bringing more false positive results.

4. Date Detection Algorithm

The Latvian NLP pipeline [9], which was used as a morphological parser of diaries to augment data with part-of-speech tags and lemmas, also contains a date recognition feature. However, due to above-mentioned irregularities in date notation techniques and the need to distinguish *metadates* from other dates in diaries, the Latvian NLP pipeline date recognition feature delivered only partially sufficient results. Therefore, it was decided to write a custom date detection algorithm, using regular expressions.

Regular expressions allow users to search a text for specific characters or sequences of characters and then perform operations on them. To account for all different cases of date_and_month placement, a general pattern was created consisting of three parts: any 0 to 25 characters at the beginning of the line (optional) + date_and_month + year (optional). Regular expressions which are the building blocks of the pattern for date detection are provided in Table 1 below. The original code was written in Javascript programming language, but is largely compatible with other programming languages.

The date detection algorithm was tested on diaries and improved until the results were satisfactory. In general, the improvements followed two lines: (1) including more regular expression patterns of *metadate* composition, when some dates were found unrecognized in processed files, and (2) narrowing down regular expression patterns to exclude false positive results. The variety of *metadate* patterns, as can be seen in Figure 2, made it impossible to imagine all combinations of symbols beforehand; therefore, such a trial-and-error method was appropriate to fine-tune the search pattern.

Table 1

Regular Expressions to Find Metadates in Text

Regular Expression	Comment
^.{0,25}/	Beginning – any 0 to 25 characters at the beginning of the paragraph.
/(19|20)\d{2}\.?\s?/	Year – a four-digit sequence, the first two digits should be 19 or 20, followed by any 2 numbers, followed by an optional dot, followed by an optional whitespace.
/\s?[0123]?\d\.?\/?\s?/	Date – a whitespace character (optional), followed by an optional 0 or 1 or 2 or 3, followed by any digit, followed by dot (optional), followed by forward slash / (optional), followed by whitespace (optional):
/jan|feb|mar|apr|mai|j[uū]n|j[uū]|aug|sep|okt|nov|dec/i	Latvian months – matches the first three letters of the month – as often authors use abbreviations and not the full month name.
/XII|XI|IX|X|VIII|VII|VI|IV|V|III|II|I[.,\s]/	Roman months – followed by a period, comma, or whitespace.
/\d{1,4}[./\\-]\s?\d{1,2}[./\\-]\s?\d{1,4}\.?/	Arab numbers – to match dd.mm.yyyy, dd.mm.yy and yyyy.dd.mm., separated by . / \ - and followed by optional whitespace.

To detect dates, the following algorithm and workflow were used. A modular, step-by-step approach helped to clearly identify what occurred in each phase, to evaluate intermediary output, to see how well the algorithm performed, and if necessary to modify it. If the date formatting of the input .txt file is clearly known, the modular approach also allows one to adjust settings focusing a regular expressions search on relevant patterns, thus decreasing false positive findings — which might be useful when working with larger files.

The .txt-in .txt-out workflow permitted to skip the building, learning, and testing of User Interface. Although, for larger-scale corpus processing, a graphical User Interface might be useful — allowing one to adjust finer settings according to input files formats and minimizing the chance of introducing errors by hand-correcting the double wrapped << >> *metadates*. Table 2 below describes the date detection algorithm step by step.

Table 2

Metadate Detection Algorithm in Five Steps

Step	Action
1	Input the text file in a .txt format.
2	The algorithm checks every paragraph of the input file. If a paragraph's beginning (first 25 characters + the following pattern) contains a date, then the beginning of a paragraph is copied, wrapped in < >, and pasted above that paragraph, with an empty line added above the wrapped text fragment.
3	In the third step, all the lines containing the wrapped < > text fragment are parsed and a *metadate* is predicted.
4	In the fourth step, < > wrapped metadates are checked for inconsistencies — unchronological dates, repeated dates, years out of legitimate years interval, etc. Suspicious dates are wrapped in double brackets << >>.
5	After the fourth step, the output file in a .txt format is returned to a humanities researcher, who manually checks all double-wrapped dates, changes them to the correct date and removes one pair of < > from corrected double-wrapped dates.

An example of the algorithm at work is given below, showing an excerpt from a sample diary with the output of every step, plus comments. The fragment is taken from the diary LFK Ak36, written by a school teacher. It describes three days at the end of 1949 in Soviet Latvia. Below is the English translation of the text which follows in Table 3 in Latvian:

17 Dec.

There is dedication in the class to improve discipline and achievements. A class behavior register has been introduced — this promotes class discipline. On the occasion of comrade Stalin's birthday on the 20th of December, pioneers Melbārdis E. and Pinka A. will take the solemn pledge. Altogether, 7 pioneers (50%) in the class will raise discipline in their achievements.

16 Dec.

Today the class carefully arranged books on the windowsill — there is no other place left — and paid more attention to the teachers' desk. I am pleased that today, for the first time in the class, I praised behavior and gave a mark of 5 for it. Comparing with the previous, one can see the uniformity of the group taking shape and the sense of responsibility setting in.

22 XII.

I have to stop again at Indulis. [..]

Table 3

Metadate Detection – Workflow Example

Step	Output	Comments
1	17. dec. Klasē vērojama centība labot disciplīnu un sekmes. Ievesta ir klases uzvedības atzīmju burtnīca - tas sekmē disciplīnu klasē. Par godu b. Staļina dzimšanas dienai 20. dec. nodos svinīgo solījumu pionieri Melbārdis E. un Pinka A. Klasē ar to 7 pionieri (50%), kas cels ari sekmēs disciplīnu. 16. dec. Klase šodien rūpīgi sakārtojusi grāmatas uz loga - citur nav vietas - un uzmanību vairāk pievērsusi arī skolotāju galdam. Priecīga. ka šodien par uzvedību pirmo reizi klasē izteicos atzinīgi un novērtēju ar 5. Salīdzinot ar iepriekšējo, var vērot sastāva vienveidības veidošanos un kolektiva atbildības sajūtu. 22 XII. Atkal jāapstājas pie Induļa. [..]	First step: input the text file. Here it is an excerpt from a diary written in 1949. / 1950.
2	<17. dec.> 17. dec. Klasē vērojama centība labot disciplīnu un sekmes. Ievesta ir klases uzvedības atzīmju burtnīca - tas sekmē disciplīnu klasē. Par godu b. Staļina dzimšanas dienai 20. dec. nodos svinīgo solījumu pionieri Melbārdis E. un Pinka A. Klasē ar to 7 pionieri (50%), kas cels ari sekmēs disciplīnu. <16. dec.> 16. dec. Klase šodien rūpīgi sakārtojusi grāmatas uz loga - citur nav vietas - un uzmanību vairāk pievērsusi arī skolotāju galdam. Priecīga. ka šodien par uzvedību pirmo reizi klasē izteicos atzinīgi un novērtēju ar 5. Salīdzinot ar iepriekšējo, var vērot sastāva vienveidības veidošanos un kolektiva atbildības sajūtu. <22 XII.> 22 XII. Atkal jāapstājas pie Induļa. [..]	The algorithm checks every paragraph of the input file. If a paragraph's beginning (first 25 characters + the following pattern) contains a date, then the beginning of a paragraph is copied, wrapped in < >, and pasted above that line and a modified copy of input file is saved to a temporary file which will be processed in the next stage. Also, an empty line is added above the wrapped fragment.
3	<17.12.1949.> 17. dec. Klasē vērojama centība labot disciplīnu un sekmes. Ievesta ir klases uzvedības atzīmju burtnīca - tas sekmē disciplīnu klasē. Par godu b. Staļina dzimšanas dienai 20. dec. nodos svinīgo solījumu pionieri Melbārdis E. un Pinka A. Klasē ar to 7 pionieri (50%), kas cels ari sekmēs disciplīnu. <16.12.1949.> 16. dec. Klase šodien rūpīgi sakārtojusi grāmatas uz loga - citur nav vietas - un uzmanību vairāk pievērsusi arī skolotāju galdam. Priecīga. ka šodien par uzvedību pirmo reizi klasē izteicos	In the third step, the temporary file is processed again, now parsing all the lines containing the wrapped < > text fragment and predicting a date from that. Here the algorithm converts from Latin and Latvian month abbreviations to Arab numbers. And the year, 1949, is correctly guessed from the incomplete date_and_month, as the year was

	atzinīgi un novērtēju ar 5. Salīdzinot ar iepriekšējo, var vērot sastāva vienveidības veidošanos un kolektiva atbildības sajūtu. <22.12.1949.> 22 XII. Atkal jāapstājas pie Induļa. [..]	given in the very beginning of the diary.
4	<17.12.1949.> 17. dec. Klasē vērojama centība labot disciplīnu un sekmes. Ievesta ir klases uzvedības atzīmju burtnīca - tas sekmē disciplīnu klasē. Par godu b. Staļina dzimšanas dienai 20. dec. nodos svinīgo solījumu pionieri Melbārdis E. un Pinka A. Klasē ar to 7 pionieri (50%), kas cels ari sekmēs disciplīnu. <<16.12.1949.>> 16. dec. Klase šodien rūpīgi sakārtojusi grāmatas uz loga - citur nav vietas - un uzmanību vairāk pievērsusi arī skolotāju galdam. Priecīga. ka šodien par uzvedību pirmo reizi klasē izteicos atzinīgi un novērtēju ar 5. Salīdzinot ar iepriekšējo, var vērot sastāva vienveidības veidošanos un kolektiva atbildības sajūtu. <22.12.1949.> 22 XII. Atkal jāapstājas pie Induļa. [..]	In the fourth step, the file is checked for inconsistencies in dates — unchronological dates, repeated dates, years out of legitimate years interval, etc. Inconsistent dates are wrapped in double brackets << >>. After the fourth step, the output file in a .txt format is returned to a humanities researcher, who manually checks all double-wrapped dates, changes them to the correct date, and removes one pair of < > from corrected double-wrapped dates.

5. Evaluation of Results and Discussion

After running the final iteration of the *metadates* detection algorithm on the corpus of 36 diaries, 15,303 *metadates* were detected, of which 456 (2.98%) were wrapped in double brackets << >> as suspicious and possibly wrong *metadates*. Upon further inspection of these 456 double-wrapped *metadates* by close reading of the texts, they were classified into following categories, see Table 4.

Evaluation of the suspicious *metadates* confirmed the importance of human evaluation.

(1) For several categories (1 and 2; also 5, 6, and 7) *metadates* were correctly formed and looked identical in the text, and only a close reading could reveal if it was a correct *metadate* or a mistake. Categories 3 and 4 (past and future events) also are correctly formed dates, but inside the narration, and only a close reading can distinguish them from *metadates*.

(2) Most often, double-wrapped <<>> *metadate*s were errors of non-chronological or impossible dates' being introduced in the earlier digitization process. However, sometimes non-chronological dates were present in the original diaries; that emphasized the need to give a unique identification number to every entry, so that both the entry's date and its sequence in the diary could be preserved when splitting data and saving separate entries.

(3) When inspecting for possible years, initially an additional check was performed to search only those years occurring within the interval of possible years as denoted in diary's metadata. However, it was noticed that sometimes authors have included later remarks in years which are outside the stated time interval of the diary (category 7).

(4) There were about a dozen *metadate* notations found in the whole corpus where date was not expressed with some variation of date_and_month but with traditional names, e.g.: Ziemassvētki (Christmas), Otrās Lieldienas (Easter Day), Vasarsvētki (Whit Sunday), Jāņi (Summer Solstice), etc. These date notations were dealt with on a case-by-case basis by humanities researchers — the reason

being that oftentimes these date notations were not precise enough to extract a specific date automatically.

Table 4
Ten categories of double wrapped << >> *metadates* found in close reading.

Nr	Description of the category	Count	In how many diaries	Comments
1.	**MULTIPLE** entries for the same day	103	24	Usually 2 entries for a day, occasionally 3.
2.	**DOUBLED** *metadate* for the same entry	4	1	In one diary, *metadate* was before and after the entry.
3.	**PAST** events	62	13	Author writes about past events, starts with a date mention; it is recognized as the *metadate* for entry.
4.	**FUTURE** events	5	5	Author writes about future events, starts with a date mention; it is recognized as the *metadate* for entry.
5.	A **TYPO** or a mistake	68	17	A typo or mistake in the day / month / year of *metadate* by author or transcriber, causing the appearance of an unchronological *metadate*.
6.	A **WRONG YEAR** by author or transcriber	7	3	For example, month in the *metadate* changes from December to January, but the year remains the same.
7.	Truly **UNCHRONOLOGICAL** sequence in a diary	81	13	Several days in mixed order, or longer sequences from another month or year, perhaps added later in the original.
8.	**WRONG** unchronological date **BEFORE** the current entry	93	15	Causes current correct date to be recognized as wrong and wrapped by << >>.
9.	**UNRECOGNIZED** start of a new **YEAR** by algorithm	4	2	After a long break (e.g., Nov 1966-Mar 1967) and no explicit year in *metadate*, algorithm marks it as a possible typo << >>.
10.	Date detection algorithm **ERRORS**	29	–	Incorrectly parsed dates with no obvious reason. Correct dates, wrapped << >> as suspicious with no obvious reason.

The cleaned *metadates* wrapped in < > further served as a separator to split a .txt file into separate day entries and save into a .json file. An example of one entry text with additional metadata is in Figure 3. Every entry has the following information: "lfk_number" — the abbreviated number of the diary in the ILFA archive; "number_of_entry_for_this_author" — allows one to detect correct non-chronological entries; "*metadate*" — a *metadate* as predicted by the date detection algorithm; "sql_date" — *metadate* transformed to .sql format for later computational analysis; "number_of_characters" — showing entry length; "when_added_to_database" — the date when added to the database.

```
  {
    "lfk_number": "LFKAk26ds",
    "number_of_entry_for_this_author": 2,
    "metadate": "<08.08.1926.>",
    "sql_date": "1926-08-08",
    "number_of_characters": 658,
    "entry_of_the_day": "8.VIII.26. Tirzas baznīcas pastāvēšanas 100 g. jubileja. Spredikoja bij. Tirzas māc. tagad
Pāvila dr. mc. Ādolfs Kundziņš, prof. V. Maldonis, Gulbenes māc. Egle un Jaunpiebalgas māc. J. Ozols, kas kā viskārs
apkalpo Tirzas draudzi. Satiku daudz vecu pazīstamu – Zēmīšu ļaudis, Rankas Ruti, Sebra Kārli u. c. Pēc dievkalpojuma
bija garīgo koncerts. Dziedāja Tirzas un Odulienes koris vecā skolotāja un ērģelnieka K. Dzelzkalna vadībā. Kā
soliste piedalījās Amanda Liberte-Rebāne. Liberts palīdzējis greznot Tirzas baznīcu. L. gimis Tirzas muižā, kur tēvs
bijis brūža pagraba meistars, māte vecā mežsarga Valviņa Rāča meita. No baznīcas mājā braucot salijām.",
    "when_added_to_database": "2021-04-21T06:54:11.686Z"
  },
```


Figure 3: One entry from the diary of Alvils Kalnietis, LFK Ak26

In total, 14,364 day entries from 36 diaries were added to the database. In further steps, data were enriched processing entries with a morphological parser of the Latvian NLP pipeline [9] and by adding demographic metadata to each entry, such as the author's gender, or author's age at the moment of writing. The division of the corpus into daily entries opens good possibilities for creating different time-related sub-corpora by combining age groups and gender, e.g. diary texts of 21–35 years old women, diaries of 36–50 years old men in the 1950s, etc.

Further computational analysis of diaries will require solving several methodological challenges, such as evaluation of the representativeness of the diary corpus. The diaries differ greatly in length, writing frequency, and stylistics, and it is yet to be determined what computational methods could offer to the general discourse of diary research, including diachronic research of single diaries and cross-comparison of different diaries according to similar properties of author's age, author's gender, and other categories. Date detection carried out on the corpus of Latvian diaries has invited new perspectives of inquiry for diaries [7], perspectives that have already been applied to periodicals [3], book printing, and other domains of time-bound written documents.

6. Acknowledgements

This paper is supported by the project "Digital Resources for Humanities: Integration and Development" (No. VPP-IZM-DH-2020/1-0001) funded by Latvian Council of Science.

7. References

[1] Dennis-Henderson, Ashley, Roughan, Matthew, Mitchell, Lewis, Tuke, Jonathan (2020). Life Still Goes on: Analysing Australian WW1 Diaries through Distant Reading. *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, International Committee on Computational Linguistics, pp. 90–104.

[2] Kiran, Kumar Bandeli, Hussain, Muhammed Nihal, Agarwal, Nitin. (2020). A Framework towards Computational Narrative Analysis on Blogs. *CEUR Workshop Proceedings,* vol. 2593, pp. 63–69.

[3] Koncar, Philipp, Fuchs, Alexandra, Hobisch, Elisabeth, Geiger, Bernhard, Scholger, Martina, Helic, Denis (2020). Text sentiment in the Age of Enlightenment: an analysis of spectator periodicals. *Applied Network Science*, No. 5(1).

[4] Lejeune, Philippe (2009). *On Diary*. Honolulu: University of Hawaii Press.

[5] Myers, Victoria, David O'Shaughnessy, and Mark Philp (eds), *The Diary of William Godwin*, (Oxford: Oxford Digital Library, 2010).

[6] Reinsone, Sanita (2020). Searching for Deeper Meanings in Cultural Heritage Crowdsourcing. In A History of Participation in Museums and Archives. Traversing Citizen Science and Citizen Humanities, Routledge's research series Museum and Heritage Studies, edited by Per Hetland, Palmyre Pierraux, Line Esborg, 2020, pp. 186–207.

[7] Reinsone, Sanita, Matulis, Haralds, Ļaksa-Timinska, Ilze. Metadatos balstīta dienasgrāmatu teksta korpusa analīze [Metadata Based Analysis of Diary Corpus]. Letonica, 2022 (to be published).

[8] Thain, Marion (2016). Perspective: Digitizing the Diary – Experiments in Queer Encoding (A Retrospective and a Prospective). Journal of Victorian Culture, No. 21 (2), pp. 226–241.

[9] Znotiņš, A. and Cīrule, E. NLP-PIPE: Latvian NLP Tool Pipeline. *Human Language Technologies - The Baltic Perspective*, IOS Press, 2018.