RussianFlu-DE: A German Corpus for a Historical Epidemic with Temporal Annotation

Tran Van Canh[1(✉)], Katja Markert[2], and Wolfgang Nejdl[1]

[1] L3S, Hannover, Germany
{ctran,nejdl}@l3s.de
[2] Institute of Computational Linguistics, Heidelberg University, Heidelberg, Germany
markert@cl.uni-heidelberg.de

Abstract. Temporally annotated corpora about historic events can be crucial to digital humanities research: they allow to extract and date events as well as reactions to them, and to construct timelines of events and of language use, among other applications. However, producing a precise corpus of a particular event in history is very challenging due to the lack of noise-free digitalized data. This paper introduces RussianFlu-DE, a temporally annotated corpus of 639 articles extracted from noisy OCR text of newspaper issues in German. All articles are about the Russian flu epidemic that took place during 1889–1893. We describe the development of RussianFlu-DE, including methods to clean different types of noise in the OCR text, and our tool for extracting Russian flu related articles. In addition, the task of temporal annotation using the TIMEX2 schema is discussed and the characteristics of the corpus compared to other corpora are presented. To show how our contribution supports epidemiology, we present some preliminary yet interesting results obtained from analyzing the articles in RussianFlu-DE. The corpus and associated tools for exploration are publicly available.

Keywords: Corpus in German · Russian flu epidemic · TIMEX2 · Temporal annotation

1 Introduction

Analyzing past events such as wars or epidemic diseases has received significant attention as knowledge obtained from such events paves the way for solving current or future similar issues. This is especially important for epidemiology due to many infectious diseases being discovered every year in many parts of the world. By analyzing historical accounts of earlier epidemics researchers can find useful insights into, for example, disease transmission methods, development stages, and community vulnerability factors. Such knowledge helps authorities prepare effective interventions, e.g., closing churches, schools, and public events, controlling methods of transport, or requiring masks to be worn in certain places

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 61–73, 2017.
DOI: 10.1007/978-3-319-67008-9_6

for reducing risks from a particular infectious disease [1,2]. However, studying any epidemic happening in the past always faces a difficult challenge because reliable datasets to conduct analysis are only rarely found. One of the main reasons is that creating a corpus of a particular epidemic event requires a lot of effort, from identifying data sources to extracting appropriate content. It is even more challenging when dealing with events that happened pre-digitization due to the lack of data and the noisy nature of the content found. Because of this, research studies on historical epidemic events are typically carried out based on documents manually collected from various repositories, with size, coverage and representativeness limited by the large amount of human effort required [3–5].

Our first goal in this work is therefore to target a historical epidemic event and create a reliable corpus for it as a main contribution for others to use. Specifically, the largest nineteenth-century epidemic of influenza, called "the Russian flu epidemic" is of our interest. The epidemic reached Europe from the East in November and December of 1889 and spread over the whole globe in the space of a few months. It was one of the first epidemics of influenza that occurred during the period of the rapid development of bacteriology. In addition, it was the first epidemic that was publicly and intensively narrated in the developing daily press, especially those published in German located in Germany and Austria [6]. However, as stated in [1], very limited information about the epidemiology of this influenza has been found in materials published in English. Motivated by these observations and by the fact that no epidemic corpus in German has been publicly available so far, we create RussianFlu-DE corpus[1], which contains 639 German news articles (stories) extracted from noisy OCR text of newspapers published during 1889–1893.

In addition, reliable extraction of interesting knowledge regarding epidemiology from a text corpus often needs to refer to the time at which reported events take place. Temporal information supports the development of techniques for timeline creation and tracking the progress of events over geographic areas. Apart from these epidemiological applications, temporal information plays an important role in many natural language processing and understanding tasks. Therefore, the extraction and normalization of temporal expressions from documents are crucial preprocessing steps in these research areas. An important application is to evaluate the quality of temporal taggers. Thus, as a second contribution, we provide a temporally annotated version of the corpus using the TIMEX2 annotation schema [7]. Finally, to show examples of how useful RussianFlu-DE is for epidemiology, we present some preliminary yet interesting results obtained from analyzing the corpus.

The remainder of the paper is structured as follows. In Sect. 2, we discuss the related work for this paper. In Sect. 3, we detail our tasks to produce RussianFlu-DE. Section 4 describes the methods used to annotate temporal expressions in the corpus. We then present in Sect. 5 some exploratory results obtained from the corpus before concluding the paper in Sect. 6.

[1] RussianFlu-DE is accessible on our project website: http://russianfluweb.l3s.uni-hannover.de.

2 Related Work

In this section, we first discuss recent studies on the Russian flu epidemic and then present relevant work for the creation of a temporally annotated corpus.

Studies on the Russian flu epidemic. As mentioned in the previous section, there is limited information about the epidemiology of the Russian flu epidemic 1889–1893. In [6], the authors conducted an analysis to examine the impact of the epidemic in 14 cities in Europe. Their results showed that the epidemic spread quickly from Saint Petersburg, Russia, to other parts of Europe with a speed of around 400 km/week and reached the American continent only 70 days after the original peak in Saint Petersburg. In addition, some detailed information about case fatality ratio and the median basic reproduction was also given. However, their work was based on reports of only two local daily newspapers in Poznań, Poland, which implies some uncertainty due to the lack of data coverage. Valleron et al. [8] presented a case study on the transmissibility and geographic spread of the Russian flu. A similar approach was followed by Valtat et al. [1] to examine the age distribution of the affected people and the mortality rate of this flu event. In a recent study, Ewing et al. [9] collected contemporary reports and explored a digital humanities approach to interpret information dissemination regarding this epidemic. The limitations common to all these studies are the heterogeneity and lack of coverage of data used.

Development of temporally annotated corpora. Since the last decade, there has been significant effort in the area of temporal annotation of text documents. Annotation standards such as TIDES TIMEX2 [7] and TimeML [10] were defined and temporal taggers e.g., DANTE [11] and HeidelTime [12], were developed. Furthermore, research challenges such as the Automatic Content Extraction (ACE) time expression and normalization (TERN) were organized where temporal taggers were evaluated. In 2010, temporal tagging was one task in the TempEval-2 challenge [13]. However, the research focus was mainly on English documents. A few temporally annotated corpora have been published, e.g., ACE EN Train corpora[2], TimeBank [14], TempEval EN [13], and WikiWars [11]. Only recently, a German WikiWars corpus consisting of Wikipedia articles in German about famous wars in history was developed [15]. Nevertheless, no historical epidemic corpus is available so far.

3 Corpus Creation

In this section, we detail the corpus creation tasks. In Sect. 3.1 we describe our methods for collecting and cleaning data. A tool for extracting Russian flu related stories from OCR text is then introduced in Sect. 3.2.

[2] See corpora LDC2005T07 and LDC2006T06 on http://www.ldc.upenn.edu.

3.1 Data Collection and Noise Reduction

Data used in this work was collected from the Austrian Newspapers Online (ANNO)[3] repository. ANNO contains almost all issues from many newspapers in Austria and Germany during the time the Russian flu epidemic took place. The data are accessible, both in scanned PDF and OCR formats. These are appropriate for our goal in terms of extracting Russian flu related stories from noisy OCR text and checking against the scanned PDF content for validity. To establish the data collection, the keywords listed in Table 1 were used to search the ANNO repository[4]. The search query was constrained to the time interval from 1889 to 1893. Empirically, these keywords are likely to appear in texts talking about diseases in general and about the Russian flu epidemic in particular, therefore resulting in a high recall collection. After preprocessing the search results we obtained 4,806 issues, which become the candidates to extract stories about the Russian flu. ANNO search always returns a whole issue of a newspaper and fully automatic extraction of individual stories is not possible.

Table 1. Keywords used to collect newspaper issues containing stories about the influenza epidemic. We aimed for high recall.

ID	Keyword	Variation	ID	Keyword	Variation
1	Influenza	Jnfluenza, Jnsolvenza	4	Grippe	
2	Epidemie		5	erkrankt	ertrankt
3	Influenza-Epidemie	Influenzaepidemie	6	Pathologie	

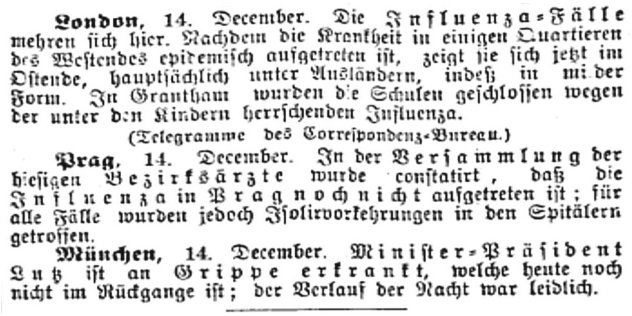
Due to the low quality of the scanned images of newspaper issues, a lot of noise is present in the corresponding OCR texts. A very frequent type of error is the so-called antistring, where the OCR output of a sequence of words consists of individual characters with spaces inbetween. For example, Fig. 1 shows a scanned image of three short messages about the Russian flu in London, Prague, and Munich, which was published on December 14, 1889 by Die Presse, and the corresponding OCR text. We can observe that a string "I f l u e n z a i n P r a g n o c h n i c h t" was produced instead of five words "Influenza in Prag noch nicht". Besides, several misrecognized words exist in the OCR text.

Our goal was to correct OCR errors but at the same time keep the language as it was so that the derived corpus pertains its historical perspective. As modern German is rather different in writing and usage of many words due to language evolution, text normalization models that were trained on modern German datasets could not be applied [16]. To cope with these issues, we adopted a snapshot of the Google-2-gram dataset for German[5] from 1885 to 1895.

[3] The Austrian National Library: http://anno.onb.ac.at.

[4] Some misspelt variations of keywords were used due to possible OCR errors.

[5] The Google-2-gram dataset is publicly available at: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.



London, 14. December. Die Influenza-Fälle mehren sich hier. Nachdem die Krankheit in einigen Quartieren des Westendes epidemisch aufgetreten ist, zeigt sie sich jetzt im Ostende, hauptsächlich unter Ausländern, indeß in mi der Form. In Grantham wurden die Schulen geschlossen wegen der unter den Kindern herrschenden Influenza.

(Telegramme des Correspondenz-Bureau.)

Prag, 14. December. In der Versammlung der hiesigen Bezirksärzte wurde constatirt, daß die Influenza in Prag noch nicht aufgetreten ist; für alle Fälle wurden jedoch Isolirvorkehrungen in den Spitälern getroffen.

München, 14. December. Minister-Präsident Lutz ist an Grippe erkrankt, welche heute noch nicht im Rückgange ist; der Verlauf der Nacht war leidlich.

London, 14. December. DieInfluenza-Fälle
mehren sich hier. Nachdem die Krankheit in einigen Quartieren
des Westendes epidemisch aufgetreten ist, zeigt sie sich jetzt im
Ostende, hauptsächlich unter Ausländern, indeß in nii der
Form. In Grantham wurden d eSchule» geschlossenwegen,
der unter den Kindern herrschenden Influenza.
(Telegramme desCorrefpondcnz-Bureau.) j
Prag, 14. December. In der Versammlung der 1
hiesigen Bezirksärzte wurde constatirt, daß die f
Influenza in Prag noch nicht aufgetreten ist; für!
alle Fälle wurden jedoch Jsolirvorkehrungen in den Spitälern l
getroffen.
München, 14. December.Minister-Präsident
Lutz ist an Grippe erkrankt, welche heute noch
nicht im Rückgänge ist; der Verlauf der Nacht war leidlich.

Fig. 1. An image of three short messages about the Russian flu in an issue published on December 14, 1889, by Die Presse and the corresponding recognized OCR text.

We used the dataset to train a bigram-based model for word segmentation (to recover words from antistrings), and for spell checking (to correct misrecognized words).

We applied our model to the 639 Russian flu stories in our corpus (see Sect. 3.2 for details on story extraction). Updates were done for 5,121 antistrings and 79,114 misrecognized words. The final corpus contains over 453,000 words. To validate our noise correction, two German native speakers proofread a random sample of 11 stories in our corpus before and after correction and identified word errors. The word error rate (WER) over all 11 texts before correction was around 18.9%, reduced to 5.5% after correction. Although our proofreaders also found some errors introduced by our model, the quality of the text has greatly improved. Table 2 shows details per story.

3.2 Tool for Extracting Russian Flu Stories

Text block classification. Manually reading thousands of newspaper issues to extract relevant stories about the Russian flu is too time consuming. In addition, a difficult challenge when extracting complete stories is that recognized OCR text blocks are very often not aligned in the same order as they were in the original image of an issue. Our approach was to automatically pre-classify OCR text blocks to identify the ones that are more likely to be part of flu-related

Table 2. Sample results of noise cleaning process. B and A prefixed columns indicate the numbers obtained before and after running our noise cleaning model, respectively.

ID	Filename	B-Errors	B-Words	B-WER	A-Errors	A-Words	A-WER
01	apr18891206.story.1	80	567	14.1%	20	573	3.4%
02	apr18891209.story.1	135	820	16.4%	50	815	6.1%
03	apr18891210.story.1	101	460	21.9%	24	403	5.9%
04	apr18891210.story.2	102	539	18.9%	33	528	6.2%
05	apr18891211.story.1	36	253	14.2%	6	236	2.5%
06	apr18891211.story.2	93	541	17.1%	30	553	5.4%
07	apr18891211.story.3	114	721	15.8%	22	721	3.0%
08	apr18891212.story.1	118	653	18.0%	31	637	4.8%
09	apr18891214.story.1	94	446	21.0%	36	443	8.1%
10	apr18891214.story.2	21	119	17.6%	6	105	5.7%
11	apr18891215.story.1	239	717	33.3%	66	726	9.0%
	Average	**103**	**531**	**18.9%**	**30**	**521**	**5.5%**

stories. For this, we adopted the KL-divergence based technique developed by Schneider [17] to build a classifier. We manually extracted and labeled 245 OCR text paragraphs, which were then used to train the model and obtained a recall of 81.5% and a precision of 68.6% when cross-validating on the training set. The output of the classifier can be used to help annotators start working on an issue by looking at suggested text blocks first, from which they then select paragraphs that are part of the same story.

Extraction tool. We implemented a Web-based tool for annotators to help build our RussianFlu-DE corpus collaboratively. The high-recall classifier described above is an underlying component of the tool that suggests to the annotators to navigate to the text paragraphs that are probably part of a Russian flu story. The main GUI of our tool is shown in Fig. 2. As an important component for the annotator, the bottom-left area displays information on text blocks derived from classifier output and the annotator's recent decisions (e.g., which text blocks were selected to complete an article). For more convenience, one just needs to click on an entry in this list to navigate to the corresponding content that is shown afterwards in the area on the right[6].

Four students worked through all 18,768 classifier-suggested text blocks of all 4,806 newspaper issues returned by the original ANNO keyword search to identify Russian-flu related stories in those blocks and surrounding blocks. The extracted 657 stories were subsequently additionally verified by another two native-speaker annotators resulting in a final 639 Russian flu stories from 42 newspapers, identified with 85.7% agreement between annotators (i.e., 18 stories

[6] Detailed guidelines for the tool are available on: http://russianfluweb.l3s.uni-hannover.de.

Fig. 2. Main GUI of our tool for extracting newspaper articles about the Russian flu from OCR text.

Table 3. List of 20 newspapers from which most of the articles about Russian flu in our corpus were extracted.

ID	Newspaper	Stories	ID	Newspaper	Sto.
01	Die Presse	72	11	Linzer Volksblatt	24
02	Neue Freie Presse	56	12	Bludenzer Anzeiger	20
03	(Linzer) Tages-Post	54	13	Vorarlberger Landes-Zeitung	17
04	Bregenzer/Vorarlberger Tagblatt	50	14	Mährisches Tagblatt	16
05	Deutsches Volksblatt	43	15	Prager Abendblatt	15
06	Wiener Zeitung	40	16	Salzburger Chronik	11
07	(Neuigkeits) Welt Blatt	38	17	Badener Bezirks-Blatt	9
08	Das Vaterland	34	18	Neue Warte am Inn	9
09	Prager Tagblatt	33	19	Volksblatt für Stadt und Land	8
10	Bukowinaer Rundschau	25	20	Teplitz-Schönauer Anzeiger	6

were removed, 548 in common, 91 in partial agreement). Table 3 shows a list of 20 newspapers from which most of the articles in the corpus were extracted. Each article was then converted into an SGML file, the format of the ACE TERN corpora containing DOC, DOCID, DOCTYPE, DATETIME, and TEXT tags. The document creation time was set to the publication date. The format complies with widely used tools for temporal annotation tasks, which we address in the next section.

4 Temporal Annotation

This section describes our work to produce a temporally annotated version of RussianFlu-DE. We first give an overview over the TIMEX2 schema (Sect. 4.1),

which we used for annotating temporal expressions. We also explain our two-stage strategy for annotating the corpus. Then, in Sect. 4.2, we present some statistics computed on RussianFlu-DE compared to other corpora.

4.1 TIMEX2 Schema and Annotation Strategy

For the annotation of temporal expressions, we followed the authors of the WikiWars corpus [11]. Particularly, we used TIMEX2 as annotation schema to annotate the temporal expressions in our corpus. The TIMEX2 annotation guidelines [7] describe how to determine the extents of temporal expressions and their normalizations. Note that, in addition to date and time expressions, such as "December 10, 1889" and "9:30 p.m.", temporal expressions describing durations and sets are to be annotated as well [15]. Examples for expressions of the types duration and set are "24 months" and "daily", respectively. The normalization of temporal expressions is based on the ISO 8601 standard for temporal information. In particular, the following five attributes can be used to normalize a temporal expression: VAL (value), MOD (modifier), SET (set identification), ANCHOR_VAL (anchor value), and ANCHOR_DIR (anchor direction).

The most important attribute of a TIMEX2 annotation is VAL, which holds the normalized value of a temporal expression. Table 4 gives values of VAL for the four examples described above. The SET attribute is used to identify set expressions. In addition, the modifier MOD is used to provide additional specifications not captured by VAL. For instance, for expressions such as "the end of December", MOD is set to END. Finally, ANCHOR_VAL and ANCHOR_DIR are used to anchor a duration to a specific date, using the value information of the date and specifying whether the duration starts or ends on this date.

Table 4. Normalized values (VAL) of temporal expressions of different types. We here assume that 9:30.p.m in the second example refers to 9:30.p.m on December 10, 1889.

Temporal expression	VAL attribute	Temporal expression	VAL attribute
December 10, 1889	1889-12-10	24 months	P24M
9:30.p.m	1889-12-10T21:30	daily	XXXX-XX-XX

Similar to Strötgen et al. [15,18], we used the Heideltime temporal tagger [12] as a first-pass annotation tool. HeidelTime is a multilingual, rule-based temporal tagger that was developed to have strict separation between the source code and the resources (rules, extraction patterns, normalization information). Because of this, HeidelTime supports several languages [12,13]. The output of HeidelTime was then imported to the annotation tool Callisto[7] for manual correction of the annotations. As in [11,15], this two-stage annotation procedure is

[7] http://callisto.mitre.org.

motivated by the fact that "annotator blindness", i.e., annotators missing temporal expressions, is reduced to a minimum. Furthermore, the annotation effort is reduced significantly since the annotator does not have to create TIMEX2 tags for the expressions already identified by the tagger. At the second stage of annotation, the stories were examined for temporal expressions missed by HeidelTime and existing HeidelTime annotations were manually corrected. This task was performed by two rounds of 2 annotators, who each worked separately on half the collection. Overall, relative to the final 7,492 temporal annotations, the annotators contributed 4.4% new temporal annotations and corrected 7.9% of HeidelTime annotations (see Table 5).

Table 5. Updates made by annotators in two rounds on the result of the HeidelTime temporal tagger. The percentage is relative to the final result.

First round of annotators		Second round of annotators	
Extraction	Normalization	Extraction	Normalization
311 (added)	561 (edited)	16 (added)	31 (edited)
4.2%	7.5%	0.2%	0.4%

Finally, the annotated files, which contain inline annotations, were transformed into the ACE APF XML format, a stand-off markup format used by the ACE evaluations. Thus, the RussianFlu-DE corpus is available in the same two formats as the WikiWars and WikiWarsDE corpora. Therefore, evaluation tools of the ACE TERN evaluations can be used with our corpus as well.

4.2 Corpus Statistics

In this section, we present some statistics regarding the length of stories and the number of temporal expressions in our RussianFlu-DE corpus compared to other corpora.

The RussianFlu-DE corpus contains 639 stories related to the Russian flu with a total of more than 453,000 tokens and 7,492 temporal expressions. In Table 6, we compare our corpus to other publicly available, temporally tagged corpora. While ACE 04 EN Train remains the largest corpus in terms of number of documents, RussianFlu-DE is the largest one regarding the number of tokens. Except for the two WikiWars copora that naturally have long narrative documents, RussianFlu-DE has significantly longer documents compared to others. RussianFlu-DE contains around 7,500 temporal expressions. Thus, the corpus is second only to the ACE 04 EN Train corpus. The density of temporal expressions indicated by Tokens/TIMEX in RussianFlu-DE is similar to the other corpora except for the two WikiWars corpora and the ACE 04 EN Train corpus. Finally, RussianFlu-DE contains 11.7 temporal expressions per document on average, which is slightly higher than the others except for the two WikiWars corpora that have an order of magnitude more temporal expressions per document.

Table 6. Statistics computed on the RussianFlu-DE corpus in comparison to other publicly available corpora.

ID	Corpus	Docs	Tokens	Tokens Doc	TIMEX	Tokens TIMEX	TIMEX Doc
1	ACE 04 EN Train	863	306,463	355.1	8,938	34.3	10.4
2	ACE 05 EN Train	599	318,785	532.1	5,469	58.3	9.1
3	TimeBank 1.2	183	78,444	428.6	1,414	55.5	7.7
4	TempEval2 EN Train	162	53,450	329.9	1,052	50.8	6.5
5	TempEval2 EN Eval	9	4,849	538.7	81	59.9	9.0
6	WikiWars	22	119,468	5,430.3	2,671	44.7	121.4
7	WikiWarsDE	22	95,604	4,345.6	2,240	42.7	101.8
8	*RussianFlu-DE*	*639*	*453,288*	*709.3*	*7,492*	*60.5*	*11.7*

5 Preliminary Exploratory Results

In addition to the RussianFlu-DE corpus, we developed associate tools and made them available so that research communities can use them to query for information and conduct explorative studies based on the corpus. We present in this section some functionalites of our tools and preliminary yet interesting results.

The corpus timeline provides statistics on the number of stories in the corpus across time and news outlet. In addition, it provides an interactive visualization

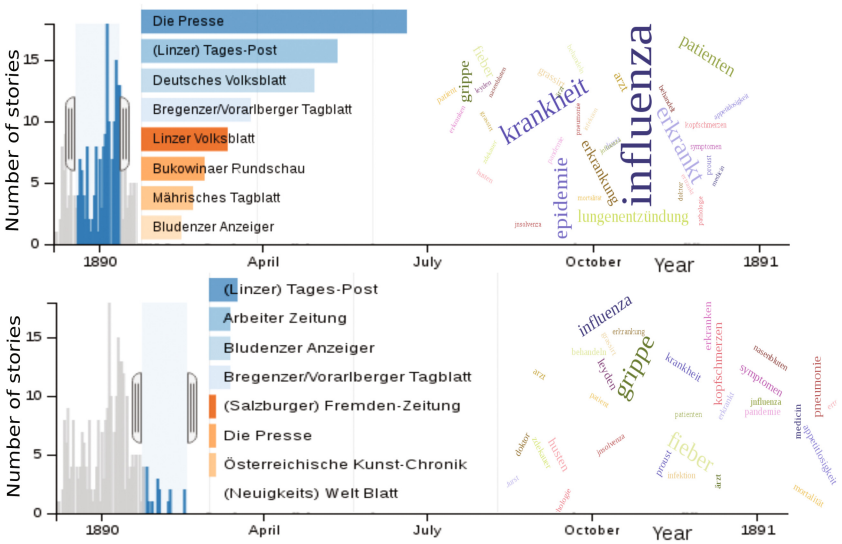


Fig. 3. Press attention on the Russian flu and topic changes over time.

Fig. 4. A pattern of frequent collocations extracted from RussianFlu-DE corpus.

from shallow semantic analysis, such as word usage and word collocations over time. As an example shown in Fig. 3, during the peak time in late December 1889 and January 1890, extensive news about the influenza was published. Newspapers were trying to narrate the outbreak as fast as possible. Words that appear significantly in the stories include *influenza, epidemic, krankheit (disease),* and *erkrankt (sick).* A short time after this peak period, i.e., in February and March 1890, fewer reports were published about the outbreak of the flu and communities started discussing its treatment more. Names of doctors appear in the news (e.g., Leyden, Proust) together with words describing symptoms such as *fieber (fever), kopfschmerzen (headache),* and *appetitlosigkeit (anorexia).*

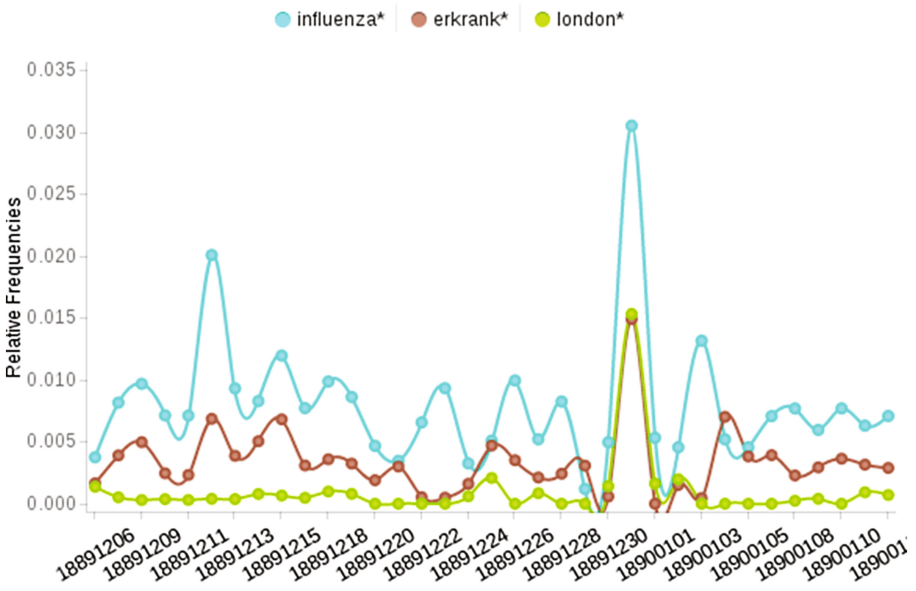


Fig. 5. High correlation between the frequencies of *influenza* and *erkrankt*, and the peak time of the flu in *London.*

Rather than such an overall view, by exploring word collocations in the whole corpus, one can find even more interesting information. For example, Fig. 4 shows a frequent pattern of word collocation describing the influenza. This pattern actually provides useful insights, both on how the media narrates the flu and the flu itself. The words *heute (today)* and *gestern (yesterday)* indicate that news about the flu was updated every day; and the word *jänner (January)* implies that the flu outbreaks happened during winter.

Figure 5 presents the co-occurrences of three words *influenza, erkrankt,* and *london* over time. It shows a strong correlation between the occurrence frequencies of *influenza* and *erkrankt*. In addition, one can observe that the peak time of the flu in London was from late December 1889 to early January 1890 as indicated in [4,19]. This suggests that the temporal distribution of terms can give us more insights into the geographical spread of the epidemic.

6 Conclusions

We have described RussianFlu-DE, a corpus of German articles about the Russian flu of 1889–1893. After discussing the methods for data collection and cleaning, we introduced our tool for extracting relevant articles from OCR text. In addition, a temporally annotated version of the corpus was produced. We further presented some interesting insights that we achieved from analyzing articles in the corpus. In future, we will (i) extend the corpus with articles from German-language newspapers published in countries other than Austria, (ii) use our temporal corpus annotation for information extraction (such as number of deaths) and (iii) investigate information spread as well as sentiment towards events by aligning articles on the same events from different newspapers.

Acknowledgments. This work is supported by the German Research Foundation (DFG) for the project "Tracking the Russian Flu in U.S. and German Medical and Popular Reports, 1889–1893" on Grant No. NE 638/13-1. We also thank you the Austrian National Library for help in data collection.

References

1. Valtat, S., Cori, A., Carrat, F., Valleron, A.J.: Age distribution of cases and deaths during the 1889 influenza pandemic. Vaccine **29**(Supplement 2), B6–B10 (2011)
2. Collinson, S., Heffernan, J.M.: Modelling the effects of media during an influenza epidemic. BMC Public Health **14**(1), 376 (2014)
3. Ewing, E.T., Gad, S., Ramakrishnan, N.: Gaining insights into epidemics by mining historical newspapers. Computer **46**(6), 68–72 (2013)
4. LeGoff, J.M.: Diffusion of influenza during the winter of 1889–1890 in Switzerland. Genus **67**(2), 77–99 (2011)
5. Yan, Q., Tang, S., Gabriele, S., Wu, J.: Media coverage and hospital notifications: correlation analysis and optimal media impact duration to manage a pandemic. J. Theor. Biol. **390**, 1–13 (2016)

6. Kempiska-Mirosawska, B., Woniak-Kosek, A.: The influenza epidemic of 1889–90 in selected european cities - a picture based on the reports of two Poznań daily newspapers from the second half of the nineteenth century. Med. Sci. Monit. **19**, 1131–1141 (2013)
7. Ferro, L., Gerber, L., Mani, I., Sundheim, B., Wilson, G.: TIDES standard for the annotation of temporal expressions. Technical report, MITRE Corporation (2005)
8. Valleron, A.J., Cori, A., Valtat, S., Meurisse, S., Carrat, F., Boëlle, P.Y.: Transmissibility and geographic spread of the 1889 influenza pandemic. Proc. Natl. Acad. Sci. **107**(19), 8778–8781 (2010)
9. Ewing, E.T., Veronica, K., Sinclair, E.N.: Look out for la grippe: using digital humanities tools to interpret information dissemination during the Russian flu, 1889–90. Med. Hist. **60**(01), 129–131 (2016)
10. James, P., Robert, K., Jessica, L., Roser, S.: Temporal and event information in natural language text. Lang. Resour. Eval. **39**(2), 123–164 (2005)
11. Mazur, P., Dale, R.: WikiWars: a new corpus for research on temporal expressions. In: Proceedings of the 2010 Conference on Empirical Methods in NLP, pp. 913–922. Association for Computational Linguistics (2010)
12. Strötgen, J., Gertz, M.: Multilingual and cross-domain temporal tagging. Lang. Resour. Eval. **47**(2), 269–298 (2013)
13. Verhagen, M., Saurí, R., Caselli, T., Pustejovsky, J.: Semeval-2010 task 13: Tempeval-2. In: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval 2010, PA, USA, pp. 57–62. Association for Computational Linguistics (2010)
14. Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al.: The TimeBank corpus. Proc. Corpus Linguist. **2003**, 647–656 (2003)
15. Strötgen, J., Gertz, M.: WikiWarsDE: a German corpus of narratives annotated with temporal expressions. In: Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2011), pp. 129–134 (2011)
16. Bollmann, M.: (Semi-)automatic normalization of historical texts using distance measures and the norma tool. In: Proceedings of ACRH-2 Workshop, pp. 3–14 (2012)
17. Schneider, K.M.: A new feature selection score for multinomial Naive Bayes text classification based on KL-divergence. In: Proceedings of the ACL 2004 (2004)
18. Strötgen, J., Armiti, A., Canh, T.V., Zell, J., Gertz, M.: Time for more languages: temporal tagging of Arabic, Italian, Spanish, and Vietnamese. ACM Trans. Asian Lang. Inf. Process. **13**(1), 1:1–1:21 (2014)
19. Honigsbaum, M.: The great dread: cultural and psychological impacts and responses to the Russian influenza in the United kingdom, 1889–1893. Soc. Hist. Med. **23**(2), 299–319 (2010)