A CORPUS LINGUISTIC CHARACTERISATION OF SPERIODIKA

Ajda PRETNAR ŽAGAR,[1]

[1]Institute for Contemporary History

The paper provides a computational analysis of sPeriodika, a historical corpus of Slovenian periodicals spanning from 1771 to 1914. The authors focus on ten prominent newspapers within the corpus, employing keyword analysis, word frequency counts, and concordance analysis to characterise the content and historical development of the Slovenian language. The study describes newspaper characteristics through computational methods, relating the findings to the post-1848 period of intense nation-building. Additionally, it addresses the challenges posed by low-quality OCR (Optical Character Recognition) in historical documents. The results are threefold: 1) a quantitative description of the selected newspapers, 2) insights into the historical progression of the Slovenian language, 3) analysis of the nature of OCR errors within the corpus. The keyword analysis reveals specific thematic orientations of the newspapers, such as agriculture, pedagogy, feuilletons, and advertising. It also underscores the newspapers' roles in nation-building. The study contributes to the field of digital humanities by demonstrating how computational tools can unlock historical insights from digitised textual data, despite the limitations of OCR technology.

Keywords: historical periodicals, keyword analysis, OCR errors, corpus linguistics

1 INTRODUCTION

The last decade saw a significant increase in academic research on historical newspaper processing (Ehrmann et al., 2023). The applications range from digitisation efforts and corpora production to computational analysis and the development of new methods.

sPeriodika (Dobranić et al., 2023) is a recently published corpus of historical Slovenian periodicals from 1771 to 1914. The corpus is extremely extensive and is based on the OCR-ed periodicals from the dLib digital library, maintained by the National and University Library of Slovenia (for the history of digital editions of periodicals, see (Eiselt, 2015)). It features some of the most important periodicals of the time, contributing to increased literacy and nation-building in Slovenia (Amon, 2008; Dović, 2006).

The paper is a corpus linguistic study as proposed in the original paper by Dobranić et al. (2024). We selected the ten most prominent news- papers, those with the highest number of publications. We provide a basic computational overview of the corpus to characterise its content. Given that the OCR quality of the corpus is low (yet on par with similar historical OCR-ed newspapers, (Kettunen & Pääkkönen, 2016)), we were interested in whether we can extract meaningful newspaper characteristics using keyword analysis, word frequencies, and concordances. The results are threefold. We provide an overall quantitative-based description of the newspapers, give insight into the historical development of the Slovenian language, and present an overview of OCR errors. By providing an overview of the corpus that would take incredibly long to complete in the absence of digitisation and annotation, we argue that annotated historical editions are extremely valuable for the Slovenian research community.

The paper is structured as follows. First, we present related work on historical newspaper analysis and the historical context of the selected journals. Second, we describe the corpus and the selected subset of ten periodicals. We characterise the newspapers with keyword analysis, which shows the specifics of each newspaper, and the list of most frequent nouns, which shows the general orientation of the newspaper. We compare the periodicals in terms of their thematic, regional, and religious orientation. Third, we critically evaluate the results and suggest potential post-processing of the published corpus based on the keyword analysis. In conclusion, we sum up the findings and present the options for future research.

2 RELATED WORK

Historical newspapers are used extensively in digital humanities, mostly due to contemporary digitisation efforts, accessible interfaces for content exploration (Ehrmann et al., 2019), and open repositories. The studies range from diachronic and comparative analyses to discourse studies, with concept shift analysis being the most prominent.

Comparative studies focus on cross-country comparisons (Mayer et al., 2022) or exploring regional differences (Park & Cordell, 2023). Diachronic studies often focus on concepts shifts (Marjanen et al., 2020; Pivovarova et al., 2019; Verheul et al., 2022), semantic change (Pedrazzini & McGillivray, 2022), or topic shift over time (Marjanen et al., 2021). Another branch of studies en-tails a more content-oriented approach, focusing on the emergence of public discourses (Marjanen et al., 2019) or nation-building vocabularies (Hengchen et al., 2021; Schoots, 2023). Some studies also focus on multilingualism (Marjanen et al., 2019; Mayer et al., 2022), a common trait of historical news-papers that makes comparative analysis particularly challenging.

Outside of digital humanities, Slovenian historical newspapers are a popular research topic. The overwhelming share of the studies focus on the nation-building processes, particularly after the 1848 March Revolution[1] (Stergar, 1977). The most comprehensive study is done by Smilja Amon, who presents an overview of Slovenian journalistic efforts (Amon, 2008). Ljubljanski zvon itself provides a great overview of the newspapers in 1885 (Anonymous, 1885). It lists 34 papers published in Slovenian, with a description, editor, publisher, and price. The final overview found 8 political papers, 3 political-economic, 4 economic, 4 religious, 4 legal, 2 pedagogical, 5 literary, 1 political-literary, and 3 humorous-satirical.

Other studies primarily focus on Kmetijske in rokodelske novice (Mihelič, 1948), which pioneered journalism in the Slovenian language.[2] Linguistic analyses are similarly popular. Only a few studies focus on content analysis and comparison. One such study is done by Štepec (1987), who analyses reporting on crime in Slovenec and Slovenski narod. Štepec ascertains that the conservative Slovenec leaves the reporting on crime primarily to the liberal Slovenski narod, as they see the reporting on crime as un-Catholic and not serv-ing any purpose. Other research on historical newspapers focused on the lan-guage question in Slovenski pravnik (Zorn, 1987), news about Istria (Marušič, 2007), fashion in women's journals (Ilich, 1999), and social-~~democrat period-icals (Kermavn~~er, 1962).

[1] The period before the 1848 March Revolution is typically referred to as pre-March or Vormärz. In the paper, we refer to the subsequent period as post-March.

[2] The first Slovenian-language periodical was Lublanske novize by Valentin Vodnik in 1797, but they were short-lived.

3 SPERIODIKA

sPeriodika (Dobranić et al., 2023) is a corpus of Slovenian historical newspapers from the 18th, 19th and 20th century. The corpus was created by Dobranić et al. (Dobranić et al., 2024). As stated by the authors, the basis are the OCR-ed data produced by different technologies in varying periods by the National and University Library of Slovenia, upon which they performed additional cleaning and preprocessing. It is available on the CLARIN.SI repository and in the noSketch Engine concordancer.

3.1 Description

There are 216 newspapers in the sPeriodika corpus with varying number of publications (max 28406, min 1). The total number of publications is 148457. As there is a significant long tail in the distribution of publications per newspaper, we decided to analyse the ten newspapers with the highest sum of publications, which represents 78% of the corpus. We decided on such metric to capture the periodicals with the largest national presence and a sufficient time span (Figure 1). Table 1 shows the ten selected newspapers with the number and share of publications (rounded to two decimal points). The papers' titles carry meaning, which broadly defines their content: Agricultural and Artisan News (Kmetijske in rokodelske novice), Slovenian holder[3] (Slovenski gospodar), Teacher's Companion[4] (Učiteljski tovariš), Slovenian Nation (Slovenski narod), Home and World (Dom in svet), The Slovenian (Slovenec), Unity (Edinost), The Ljubljana Bell (Ljubljanski zvon), (Kinder)garten (Vertec), Soča.[5]

3.2 Keyword comparison

We used noSketch Engine to extract keywords for all ten periodicals. We compared them to the entire corpus, meaning we extracted words (lemmas as extracted by noSketch) that are highly represented and thus statistically significant for a given subcorpus. Lemmatization was done with the CLASSLA-Stanza

[3] Gospodar can mean a holder, a lord, a master.

[4] Tovariš can mean a companion or a comrade. The newspaper became related to politics only after 1900.

[5] Soča is a river in Western Slovenia.

Newspaper	no. of publications	% publications	no. of tokens
Kmetijske in rokodelske novice (KRN)	28406	19	29,834,568
Slovenski gospodar (SG)	16009	11	22,602,374
Učiteljski tovariš (UT)	15674	11	24,337,225
Slovenski narod (SN)	14039	9	183,294,799
Dom in svet (Ljubljana) (DS)	11073	7	32,326,449
Slovenec (1873) (SVN)	10897	7	137,506,802
Edinost (Trst) (ED)	8371	6	98,274,429
Ljubljanski zvon (LZ)	3923	3	15,590,800
Vertec (1871) (VT)	3515	2	3,170,465
Soča (SČ)	3367	2	38,879,707

Table 1: Newspapers with the highest number of publications in the sPeriodika corpus.

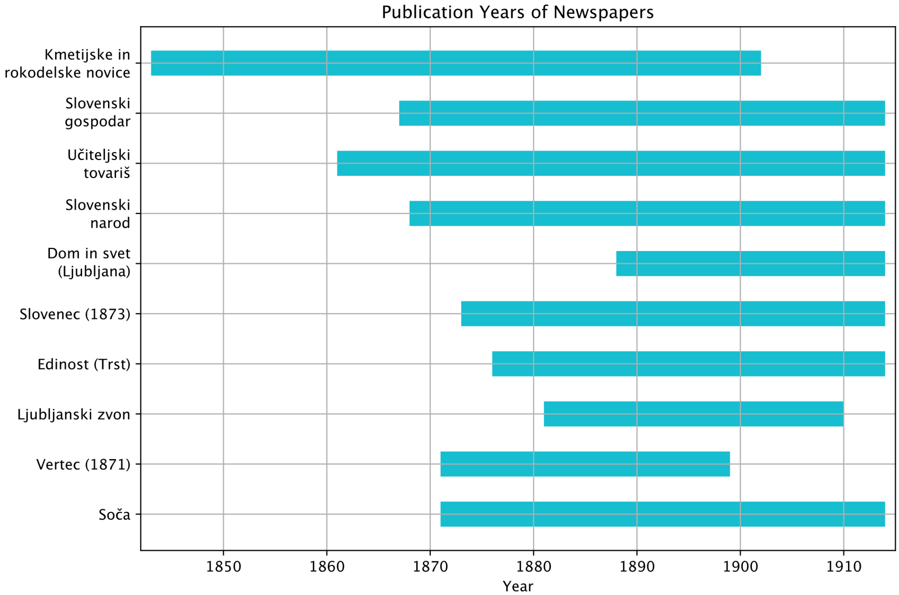


Figure 1: Publication years for the ten selected periodicals.

pipeline, as reported in the original sPeriodika paper (Dobranić et al., 2024). The keyness score, upon which the keywords are identified in noSketch, is computed with a simple maths method (Kilgarriff, 2009) and a smoothing parameter N=1 (default).

The formula for the keyness score, also known as ”simple maths” (Kilgarriff, 2009), is as follows:

$$\frac{fpm_{rmfoucs}+N}{fpm_{rmref}+N}$$

where $fpm_{rmfoucs}$ is the normalised (per million) frequency of the word in the focus corpus, fpm_{rmref} is the normalised (per million) frequency of the word in the reference corpus, and N is the smoothing parameter.

We analyse the top hundred words and present the first ten in Table 2. We omit the obvious OCR errors because we want to demonstrate the key content of the periodical, not the accidental errors. We report the number of OCR errors (per cent of errors in 100 hits) in the final row.

Kmetijske in rokodelske novice is true to its name. It discusses agricultural topics (kmetovavec, žlahen,[6] žebec[7]) and regional news (Kranjska). It was the first full-fledged Slovenian language newspaper, and as such, it contains certain archaic words more than the other newspapers (onidan, en malo). The remaining words cover diverse categories, from newspaper sections (novičar) and finances (dnar) to news on Russia (rusovski) and national enlightenment topics (čitavnica).[8] Keyword analysis testifies to the wide variety of topics the newspaper covered and its longstanding central role in the cultural life of Slovenians in that period (Stergar, 1977).

Slovenski gospodar is the first paper in the list heavily affected by errors in OCR (94 %). The 94 % error rate refers to the keyword analysis results, not the entire periodical content. Inspection in a concordancer reveals that, typically, the letter n is transcribed as a (sloveaski –> slovenski, aaš –> naš, aemški –> nemški) and v as 7 (pra7). Other keywords reveal that mistaking č for 6 is also common. There are also mentions of Stajerc, which is a misliteration of Štajerc.

[6] Žlahen means noble and refers to different breeds, from cattle and bulls to fruit trees.

[7] Žebec is an archaic word for žrebec and means stallion.

[8] ”Ćitavnica” was more frequent in the earlier editions of KRN, where it was later substituted with ”čitalnica”.

It sometimes means a person from the Styria region, but most commonly, it refers to the Štajerc periodical, published between 1900 and 1918. The tone is very derogatory since Slovenski gospodar was a pro-Catholic and conservative periodical, while Štajerc was a more progressive pro-German newspaper (extensively described in (Jezernik, 2022). The few relevant keywords refer to fairs (sermon), happen (izgoditi), golden coin (fl), school union (šulverein), people (Dr Franc Radaj, MP; Franc Kosar), esteemed (vlč, velečastiti), and posilinemec (a mocking expression for pro-German Slovenians).

Učiteljski tovariš is also true to its name. Most keywords refer to pedagogy (zavezin,[9] konvikt,[10] učiteljstvo, učiteljski, lehrerbund, pedagoški,koleginja, ljudski). There is a political aspect to the debate with continuous mentions of ”Slomškar”, which refers to the competing ”Slomšek Union”, a union of Catholic teachers. As for ”tovarišica” (comrade, colleague, teacher), it is unclear whether the word has a political connotation or not, even from collocations. However, the two references to female colleagues (tovarišica and koleginja) are highly represented in Učiteljski tovariš, indicating the periodical perhaps treated female colleagues with a higher degree of equality. The periodical does have a much higher frequency of mentions of the two words relative to the general corpus. However, collocations do not reveal any special differences in context. Učiteljski tovariš also has a high degree of German loanwords (Lehrerbund, Lehrer, Volkschule, Lehrerschaft, Gesuche, Vorgeschriebenen) and mentions of people (Črnagoj,[11] Jelenc, Maier, Strmšek, Režek, Požegar, Gangl).

Keyword analysis of Slovenski narod reveals many specific sections from the newspaper. The newspaper regularly published train schedules for Austrian railways (amstetten, pontabel, selzthal), reports from the Vienna stock exchange (prior oblig.), meteorological reports (wind directions), and specific advertisements (Moll Seidlitz powder, Revaliescere du Barry, Berger Kotran soap). Some words refer to the leading paragraph in the paper, which gave instructions for submissions to the paper (izvoti,[12] četiristopne). There are a few OCR errors characteristic of the Slovenski narod, perhaps due to the font choice

[9] Zaveza refers to the Association of Austrian Yugoslav Teaching Unions.
[10] Konvikt is an educational facility with a full board, mostly for priests.
[11] Fran Črnagoj was a teacher and businessman.
[12] This is a wrong lemma of ’izvole’, which means ’should it please them’.

(tuđi, tuđ,[13] ćel[14]) Some of the results might be the result of over-correction, as Dobranić et al. (Dobranić et al., 2024) report statistic-based joining of split words (Trammwaydrušt, Stražatoplice).

Dom in svet (Ljubljana) is heavily literary and art-oriented. The periodical is characterised by the names of fictional characters (bodriški nadknez Gotšalk, Viljenica, Virida, Maruška, Ančka) or authors (Podgoričan) of stories the paper continually published. Much of their news mentions art pieces (spominiki, bilina, pasionski) and references publications (a text on cuneiform memorials, written by F. Sedej and published in the same newspaper). The most surprising is the heavy influence of the Slavic art world on the paper. Dom in svet regularly writes biographies of Central, Eastern, and Southern Slavic authors, and lists Slavic publications (especially in Russian, Serbian, and Croatian).

Similarly to Slovenski narod, keyword analysis of Slovenec revels specific sections of the newspaper, for example, reports from the Vienna stock exchange (vravnaven, salmov, dunavski, napoleondor, napoleond,[15], waldsteinov), meteorology report, and a feuilleton Pismo Boltatovega Pepeta,[16] written in a dialect (gespud, tku, kokr). There are some recurring advertisements, for example, for the Mercur Exchange Limited Company (kurzen), glass-making workshop, and an oil paint store. Several keywords refer to South-Eastern Europe (Croatia, Hungary, Bulgaria), slightly denoting the political orientation of the periodical. However, we expected a much higher ratio of political keywords due to the newspaper's importance in the Slovenian political space. Many keywords stem from the newspaper's header, where practical information on the subscription and distribution of the periodical was given. However, other periodicals, such as Slovenski narod, Slovenski gospodar, Edinost and Soča, also had a substantial header. The high prevalence of header words is perhaps due to the linguistic specifics of Slovenec's header.

Edinost (Trst), a paper published by the Slovenes in Italy, specifically Trieste, contains many marketing-related words. Many of them refer to streets or locations of business (barriera, nuova, vecchia, piazza, galatti), specifically, 68 %. Most are Italian street names, but there are also mentions of Istrian towns

[13] Both versions of tudi, meaning also.
[14] Correctly celo or čelo, meaning even or forehead.
[15] Both terms are literal transcription of napoléon d'or, a gold coin from France.
[16] A pseudonym for Srečko Magolič (Steska & Stelè, 2013).

(Pula, Rovinj). Edinost covered the Istrian region until 1902 when the Political Society of Croats and Slovenians in Istria was formed (Darovec, 2023, 66). When mentioning the Primorska region, mentions relate mostly to the weather forecast and the journal's subtitle (Glasilo političnega društva "Edinost" za Primorsko). There are also mentions of currencies (nvč is an abbreviation for "novčič", a coin at 1/100 of "goldinar") and advertisement space (inseratni refers to the newspaper's department for advertisement). Advertisement space is characterised by recurring ads for coffee (kava Santos good average), health services (izdiranje, plombiranje, ambulatorij), and food items (pekarna, butejka). Like other periodicals of the time, Edinost regularly published train schedules. "Medpostaja" and "Pula" are mostly used in the context of railway schedules, similar to Slovenski narod, but focused on Italian railways. Railway schedule news items show that the paper is very practical; it offers advertising space for local businesses and gives information on transportation. Many periodicals of the time had similar information (e.g. Slovenski narod).

Ljubljanski zvon was the leading literary publication of the time. Most of the top ten keywords contain references to literary characters (gojko, samorad, trenk, abadon, zdenka.). 29 % of keyword results are character names, highlighting the literary nature of the periodical. However, not all content was fictional. There are references to Slovniški razgovori (Grammatical discussions), where the periodical published lectures on proper Slovenian spelling and grammar (sedanjik, sgl, miklosich, dovršnik), and Štrekelj's Jezikoslovne mrvice (Linguistic nuggets), where the author was explaining the grammatical composition, meaning, and origin of certain words (subst). Many keywords are OCR errors, namely 36 %. The keyword issue with Ljubljanski zvon is somewhat particular. It is not only that, similarly to Slovenski gospodar, the top keywords are incorrectly transcribed (OCR-ed) words. The errors are linked intimately to the literary nature of the periodical. It is the only periodical selected for analysis that consistently uses diacritics on vowels. Diacritics are uncommon in Slovenian, but in the specific newspaper, it was likely used to stress the rhythm and correct pronunciation of the word. However, this stylistic choice causes issues for the OCR model.

Vertec (1871) contains many stories and is, thus, similar to Dom in svet and Ljubljanski zvon, characterised by literary characters (Marijca, Marijec,[17] Katarinka, Ivanek). The ratio of literary character mentions in keyword results is 38 %. Unlike in other periodicals, the names are predominantly diminutives, reflecting the newspaper's orientation towards the youth. However, sometimes, a name refers not to a literary character but to a real person. The periodical listed the authors of correct solutions for its puzzles by name and location. Other keywords are bucolic, family- or nature-oriented (dedek, sestrica, ptičica, čmrlj, lisica). OCR error rate for this periodical is fairly high, at 36 %.

Soča published several translated works, including Alexandre Dumas' Three Musketeers (Athos, Porthos, Artagnan, Aramis) and Count Monte Cristo (Villefort), Henryk Sienkiewicz's Quo Vadis? (Vinicij) and The Knights of the Cross (Zbišek), and Maxim Gorky's Foma Gordeyev. The keywords, in total, include 23 % of character names. There are some regional specialities in the newspaper, for example, the word ”nunc”, which in the Gorizia dialect refers to an older familiar man. The regional character is also reflected in the mentions of local political figures, such as Alojzij Pajer-Monriva, a pro-Italian lawyer and politician, and Ivan Berbuč, a political and co-editor of Soča. A fun finding is the keyword ”prismojenec”.[18] ”Prismojenec” is a nickname for Primorski list, a conservative periodical standing in opposition to Soča, similar to how Slovenski gospodar stood in opposition to Štajerc. Soča, on the other hand, was contentwise more similar to Slovenec (Marušić, 2005, 326). The periodical contains 53 % OCR errors, making it one of the most difficult periodicals to analyse. A typical OCR error for this specific periodical is the omittance of the caron (uze,[19] dezelni, drzaven, goriski, u2e). Moreover, the periodical has low-quality images, making OCR errors even more likely.

3.3 Characterisation by nouns

To further characterise the periodicals, we retrieved lists of the most frequent nouns for each journal from the noSketch Engine 3. While keywords describe the particularities of each journal compared to the entire corpus, they are often skewed towards OCR errors and coincidental recurring feuilleton full of literary

[17] Marijec is an erroneous lemma for the word Marijca

[18] Prismojenec in Slovenian means a wacko.

[19] Originally uže.

characters. To better understand the general nature of each journal, we observed the most frequent nouns. We decided on nouns to avoid having too many stopwords in the results. Nouns are, generally, a good indicator of the content.

It turned out, nouns were not very informative. All periodicals contain words pertaining to the journalistic form, i.e. dates and place names. Some results are OCR errors, which are the most frequent in Slovenski narod, Edinost, and Soča. It is expected the OCR error rate to drop, since we only asked for CLASSLA-identified nouns. However, Edinost was the only periodical where the error rate increased significantly. The newspaper is so highly characterised by place names and advertisements, that they overtook OCR errors in keyword analysis results. The errors in noun results are comparable to other periodicals. Errors aside, the most frequent 100 nouns reveal a general orientation of each newspaper.

Kmetijske in rokodelske novice prominently features the words country (dežela) and city (mesto), showing the newspaper's focus on the city and countryside relations. It contains references to politics (zbor, vlada, odbor, poslanec) and to national identity (narod, beseda, jezik). Due to its popularity, the periodical became a central publication for the Slovenian national movement (Dović, 2023). Slovenski gospodar similarly references politics (zbor, društvo, poslanec, volitev, okraj) and city-countryside relations while also showing its religious orientation (cerkev, nedelja). Učiteljski tovariš is highly focused on pedagogy (šola, učitelj, učiteljstvo, otrok, učiteljica, knjiga, učenec), with some organisational words inbetween (društvo, svet, zbor, odbor). Vertec is also domain-specific, generally focusing on family (mati, oče, otrok) and storytelling (človek, bog, čas, mesto, hiša). Slovenski narod, while littered with single letter ”nouns” (i.e. errors), appears as a mostly political periodical (narod, zbor, vlada, Dunaj, Slovenec, stranka). Slovenec is another politically oriented periodical (vlada, mesto, društvo, zbor, narod). Dom in svet is more contemplative, mostly discussing the position of the man in the world (human, time, work, life, world, heart), with literary (knjiga, pisatelj, pesem, jezik) and religious emphasis (cerkev, bog, duša). Ljubljanski zvon is similarly contemplative but with much greater literary emphasis (knjiga, človek, beseda, življenje, delo, jezik, srce, narod) and a lack of religious themes. Regional identity is stressed in Edinost (Trieste being highly ranked) and Soča (Gorizia). Both periodicals show political/national orientation (društvo, narod, vlada).

Kmetijske in rokodelske novice, Slovenski narod, Učiteljski tovariš, Slovenec, and Ljubljanski zvon also frequently mention Ljubljana, showing their central geographical orientation. Conversely, Slovenski gospodar mentions Maribor instead, revealing its focus on the Styrian readers.

Strong nation-building forces in Slovenia defined the late 19th and early 20th centuries. The media landscape of the time greatly contributed to forming and expanding ideas of national identity, Slovenian culture, and political emancipation (Amon, 2008). Digitised editions of historical papers enable observing, comparing, and quantifying nation-building discourses. Below, we provide a quick glimpse into two nation-building aspects of Slovenian historical newspapers: the emergence of the ethnonym Slovenian and a comparison of post-March revolution discourses in Kmetijske in rokodelske novice.

7 of 10 periodicals have the word Slovenian (Slovenec) among the top 30 most frequent nouns. Upon inspection, the word Slovenian appeared only after 1843 when Bleiweis's Kmetijske in rokodelske novice was first published. The lack of the ethnonym ”Slovenian” before 1843 can be partially attributed to many periodicals before this year being in German due to the strict pre-March censorship (Dović, 2006). However, as Dović argues (Dović, 2023), Novice pioneered the ethnonym Slovenia and Slovenians into Slovenian periodicals of the post-March period.

We examined collocations for the word Slovenec to determine whether the mentions mostly refer to the periodical Slovenec or the ethnonym. The most frequent collocates are Croats, Carinthian, and Trieste.[20] References to the periodical come in at fourth place, where the collocation is the quotation mark. Thus, mentions of the periodical appear in quotations, while the ethnonym appears without them.

[20] as an adjective for the Slovenes in Italy.

Table 2: Top 10 keywords (lemmas) in selected periodicals. The cells contain a lemma and its frequency in the given periodical. The final row reports a percentage of OCR errors in top 100 keywords.

Rank	**KRN**	**SG**	**UT**	**SN**	**DS**	**SVN**	**ED**	**LZ**	**VT**	**SČ**
1	unidan (1,552)	sejmov (843)	zavezin (2,265)	amstetten (11,058)	nadknez (738)	vravnaven (3,299)	nvč (12,057)	gojko (889)	marijca (269)	athos (2040)
2	novičar (3421)	izgoditi (481)	konvikt (5,486)	izvoti (7,416)	virida (798)	gespud (3,447)	galatti (5,504)	samorad (679)	otiti (475)	porthos (1,411)
3	čitavnica (2,044)	fl (12,467)	učiteljstvo (54,905)	pontabel (6,225)	spominik (1,029)	tku (4,680)	barriera (7,162)	trenk (713)	štir (368)	artagnan (1,369)
4	rusovski (1,714)	šulverein (677)	učiteljski (58,083)	selzthal (8,551)	bodriški (631)	salmov (2,996)	inseraten (7,641)	abadon (549)	vrtčev (220)	aramis (1,253)
5	kmetovavec (2,481)	radaj (541)	slomškar (1,244)	oblig (6,752)	viljenica (638)	kokr (3,535)	nuova (7,977)	zdenka (826)	katarinka (172)	nunec (1,946)
6	dnar (2,238)	vlč (903)	tovarišica (4,632)	franzensfeste (7,256)	juriš (912)	napoleondor (3,206)	konsorcija (5,091)	groga (1,046)	ivanek (181)	zbišek (1,004)
7	žlahen (1,433)	kosar (673)	koleginja (1,031)	četiristopen (3,690)	gotšalk (610)	kursen (2,771)	pula (7,343)	cetinovič (334)	pesenca (203)	meljavec (928)
8	krajnski (3,076)	posilinemec (463)	lehrerbund (902)	steyr (5,488)	maruška (670)	dunavski (4,189)	vecchia (6,292)	dramatiški (642)	marijec (155)	villefort (846)
9	žebec (632)	-	pedagoški (2,796)	osoben (28,671)	podgoričan (996)	waldsteinov (2,349)	medpostaja (3,331)	obsezati (1,943)	vzpomlad (176)	vinicij (821)
10	enmalo (823)	-	črnagoj (779)	vara (13,567)	ančka (1,407)	napoleond (2,234)	piazza (14,364)	premec (381)	ivanko (170)	foma (916)
errors	5%	92%	12%	19%	1%	15%	0%	36%	36%	53%

Table 3: Top 10 keywords (nouns) in selected periodicals. The cells contain a noun and its frequency in the given periodical. The final row reports a percentage of OCR errors in top 100 keywords.

Rank	KRN	SG	UT	SN	DS	SVN	ED	LZ	VT	SČ
1	leto (82,258)	dan (69056)	šola (114,848)	leto (443,770)	leto (59,446)	leto (343,343)	Trst (258,913)	leto (39,468)	mati (6,266)	Gorica (80,144)
2	dan (68,933)	leto (51,176)	učitelj (97,051)	dan (408,203)	dan (48,748)	dan (321,052)	dan (216,971)	dan (27,830)	dan (5,835)	dan (75,524)
3	zbor (39,665)	zbor (28,349)	leto (67,049)	Ljubljana (302,647)	človek (42,447)	Ljubljana (229,561)	leto (157,861)	knjiga (20,532)	leto (5,744)	leto (69,817)
4	čas (37,683)	društvo (27,645)	dan (58,550)	ura (237,529)	čas (37,518)	ura (167,718)	ulica (140,912)	čas (19,748)	oče (4,515)	društvo (42,044)
5	človek (32,521)	poslanec (23,641)	učiteljstvo (54,905)	mesto (192,520)	delo (35,653)	vlada (153,937)	ura (130,535)	človek (16,778)	otrok (4,499)	zbor (41,046)
6	Ljubljana (30,918)	Slovenec (23,277)	društvo (50,566)	društvo (185,375)	življenje (34,640)	mesto (148,402)	društvo (124,234)	gospod (15,485)	človek (4,279)	Slovenec (35,437)
7	dežela (30,825)	kmet (23,255)	svet (40,158)	narod (181,277)	svet (33,444)	društvo (141,115)	cena (104,929)	mesto (14,896)	bog (3,788)	mesto (34,216)
8	mesto (30,551)	šola (23,213)	otrok (36,984)	zbor (166,316)	knjiga (33,096)	zbor (140,109)	mesto (91,023)	beseda (14,832)	čas (3,334)	čas (33,146)
9	šola (30,037)	človek (21,664)	Ljubljana (30,369)	vlada (164,447)	srce (31,636)	čas (134,440)	vlada (83,082)	nega (12,902)	gospod (3,060)	ura (32,886)
10	kraj (28,716)	mesto (21,129)	čas (29,185)	čas (163,074)	mesto (30,324)	narod (131,146)	ulica (81,713)	pisatelj (12,759)	roka (2,939)	gospod (31,052)
errors	5%	17%	13%	21%	3%	15%	22%	7%	18%	20%

4 DISCUSSION

We characterised the periodicals using two keyword analysis approaches, one with lemmas and one with nouns. This portrays a landscape of periodical at the turn of the 19th century, and supplements previous manual analysis of Slovenian newspapers. Periodicals are typically characterised by their proclaimed focus (KRN, Učiteljski tovariš), feuilletons and advertisements (Dom in svet, Slovenski narod itd.), or, alas, their OCR errors (Slovenski gospodar).

The identified importance of feuilletons and advertisements aligns with the previous research on historical Slovenian periodicals. Feuilletons, a part of a newspaper devoted to fiction, played an important role in the development of the Slovenian prose (Dović, 2006). Feuilletons were the first public venue for Slovenian authors to publish their work and reach a wider audience. Of course, keyword analysis only pointed to specific literary characters, which is expected as the technique determines words that appear uniquely in the subset. Thus, one cannot say that keyword analysis pinpointed the importance of feuilleton – the discovery was incidental.

On the other hand, the importance of advertising space was better characterised by the method. The ratio of editorial to advertisement space was at 4:1 in the late 19th century (Dović, 2006), making advertising space an extremely relevant part of the newspaper. While some keywords point to specific advertisers, they also point to the general advertising language (inseraten, nvč).

Keyword analysis reveals that the periodicals were published when the standard Slovenian language was still being formed. Many papers are characterised by their specific writing of standard Slovenian words. Almost every periodical has a set of words that characterise their approach to Slovenian spelling. For example, Kmetijske in rokoldelske novice writes nograd for vinograd (vineyard) and berž for brž (as soon as). Slovenski narod writes denes for danes (today) and sklenica for steklenica (bottle). Edinost writes menenje for mnenje (opinion), zvršetek for konec (end), and žnjo/žnjimi for z njo/z njimi (with her, with them). Vertec writes otiti for oditi (leave), vzpomlad for spomladi (in spring), and rekši for je rekel/rekla (said). Even Ljubljanski zvon, which was at the forefront of gramamtical efforts at the time, contains words that are considered archaic in modern Slovenian, i.e. obsezati for obsegati (to cover), zanimljiv for zanimiv (interesting), smijati for smejati (to laugh).

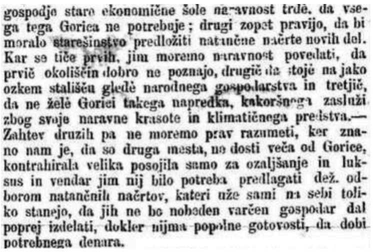
Finally, some periodicals have too many OCR errors to properly characterise them by content (Slovenski gospodar, partially Soča). OCR errors were calculated on keyword analysis results, which provide the 100 most characteristic keywords for a given subcorpus. Out of these, we manually labelled OCR errors and summed them. We considered a lack of carons an OCR error since the word without carons is counted as distinct from the word with a caron (drzaven vs državen) or can mean a different word altogether (čelo/celo). Percentages are reported for total OCR results. There were, in total, 1000 keyword results, which contained 266 errors. Note that periodicals were digitised using different OCR models, thus leading to periodical-specific errors.

Some OCR errors are recurring and reflect an underlying weakness of OCR models. The most common errors (24 %) are the letters n, s, or š transcribed as a. The errors are most common in Slovenski gopodar, the periodical most affected by OCR errors. The second most common error (21 %) is the lack of diacritics (stajerski, drzaven), while in the third place (9 %) are diacritics transcribed as numbers, specifically 6, 7 or 2 (dom6v, rek6, už6, u2e, pra7). Diacritics are often also transcribed as d (takdj). The preceding letter n often means the word begins with quotation marks (nkaj, nne, njaz). Conflation is very common, with č and e (oee, užč), i and l (ijubi, nefranklran), c and e (Marijea, evetice), and u and n (nčenki) conflated.

Diacritics and carons pose a particular problem in the transcription of sPeriodika. Here is an example from Ljubljanski zvon, the only periodical that regularly uses diacritics on vowels to denote word stress (Vertec uses them occasionally):

1. *Takó kričálo vse je gôri náme.* (original)
2. *Takd kričdlo vse je g6ri ndme.* (transcript)
3. *’ma krimu’ vse ie gori ”Am,* (tesseract)

The mistakes visually make sense. ó and á are transcribed as d (or occasionally 6), ô as 6, á also as ä, é as č. Nevertheless, issues with transcription limit the semantic analysis of significant keywords.



gospodje stare ekonomične šole naravnost trdė, da vsega tega Gorica ne potrebuje; drugi zopet pravijo, da bi moralo starešinstvo predložiti natančne načrte novih del. Kar se tiče prvih, jim moremo naravnost povedati, da prvič okoliščin dobro ne poznajo, drugič da stojė na jako ozkem stališču gledė narodnega gospodarstva in tretjič, da ne želė Gorici takega napredka, kakoršnega zasluži zbog svoje naravne krasote in klimatičnega predstva.— Zahtev druzih pa ne moremo prav razumeti, ker znano nam je, da so druga mesta, ne dosti veča od Gorice, kontrahirala velika posojila samo za ozaljšanje in luksus in vendar jim nij bilo potreba predlagati dež. odborom natančnih načrtov, kateri uže sami na sebi toliko stanejo, da jih ne bo nobeden varčen gospodar dal poprej izdelati, dokler nijma popolne gotovosti, da dobi potrebnega denara.

Figure 2: A section of the Soča newspaper, with a poor scan quality.

Vertec has specific OCR errors. While not exclusive to this periodical, Vertec suffers strongly from character conflation. Characters and character sets m, u, and ru are often wrongly transcribed. The errors stem from character (set) similarity; thus, m is transcribed as ra, ni, or in. U is transcribed as ii, and ru as ni or m. V is often transcribed as r, ó as d or 6.

Slovenec and Edinost have a different problem. In Slovenec, 29 % of keywords refer to the newspaper's header. In Edinost, 68 % of keywords refer to Italian street names. These results do not tell much about the content besides a heavy influence of specific periodical sections. There are better techniques than keyword analysis to determine the journal's content in both cases.

For periodicals with frequent errors in the top keywords, we compared the frequencies of wrongly ORC-ed words to their original form. The erroneous "sloveaski" appears 1,855 times in the corpus, while the correct version "slovenski" appears 45,759. The top keywords cannot be analysed semantically, as all occurrences of the wrong word should be first converted to the correct form. However, the error is significantly more frequent in Slovenski gospodar compared to any other periodical. The discrepancy in frequencies means the error word characterises this particular journal and could be used as a part of post-processing. In other words, such erroneous words could be subsequently corrected in the selected publication. As Strange et al. demonstrate (Strange et al., 2014), OCR correction can be crucial for certain text analysis techniques, such as keyword analysis (and less for others).

Alternatively, the error rate could indicate candidate journals for re-scanning. Certain scans are already of poor quality or were among the first periodicals OCR-ed. Contemporary OCR solutions could provide a much better result than the existing version. However, scanning the entire corpus again is tedious and unnecessary. Thus, having a list of candidate periodicals for re-scanning is a good way forward. Based on our results, Slovenski gospodar and Soča (2) would benefit from both re-scanning and a modern OCR transcript, while Ljubljanski zvon would benefit only from an OCR (scans are already good).

However, contemporary state-of-the-art large language models (LLMs) can already achieve incredible transcription accuracy. Here is a GPT4-o transcription of Figure 2: *"gospodo staro ekonomične šole nezavnost trde, da vsega tega Gorica ne potrebuje; drugi zopet pravijo, da bi moralo starešinstvo predložiti natčene načrte novih del. Kar se tiče prvih, jim moramo naravnost povedati, da prvič okolišin dobro ne poznajo, drugič da stojé na jako ozkem stališču glede narodnega gospodarstva in tretjič, da ne želé Gorici takega napredka, kakoršnega zasluži zaradi svoje naravne krasote in klimatičnega prečistva. Zahtev drugih pa ne moremo prav razumeti, kar znano nam je, da so druga mesta, no dosti veča od Gorice, kontrahirala velika posojila samo za ozaljšanje in luksus in vendar jim ni bilo potrebno predlagati dež. odboru natancnih načrtov, kateri že sami na sebi toliko stanjo, da jih ne bo nobeden varčen gospodar dal poprej izdelati, dokler njim popolne gotovosti, da dobi potrebnega denarja."*

The capabilities of LLMs and large multimodal models (LMMs) can overcome poor scan quality almost out-of-the-box. They outperform modern OCR solutions in post-OCR correction (Thomas et al., 2024) and in direct OCR (Liu et al., 2024), even for complex compositions such as old Chinese newspaper clippings (Chow, 2024). This opens up exciting venues for historical research (Garcia & Weilbach, 2023), especially in addressing corpora quality (OCR), but also for content summaries, event detection, trend analysis, and semantic search.

5 CONCLUSION

Keyword analysis reveals several aspects of the periodicals. Some papers are characterised by their general content, such as agriculture (Kmetijske in rokodelske novice) or pedagogy (Učiteljski tovariš). Some are characterised by the recurring feuilletons they publish (Dom in svet, Slovenec, Vertec, Soča). Others still are characterised by their advertising space (Slovenski Narod, Edinost). Slovenski gospodar, unfortunately, contains too many OCR errors for keyword analysis to reveal meaningful insights. Consistent OCR errors in the periodicals could be addressed in post-processing.

We substantiated the results with the most frequent nouns to alleviate the issues with keyword analysis. Many newspapers of the time focused on Slovenian nation-building, either through reports on (inter)national relations, discussions of politics, or debates on the role of the language (Kmetijske in rokodelske novice, Slovenski gospodar, Slovenski narod, Slovenec, Edinost and Soča)Among these, city vs countryside relations are a prominent topic (Kmetijske in rokodelske novice, Slovenski gospodar). Učiteljski tovariš and Vertec are domain-specific, rarely discussing topics outside their proclaimed focus. Slovenski gospodar and Dom in svet reveal the highest religious orientation. However, many periodicals of the time discussed the role of religion in nation-building.

The computational overview provides several opportunities for further analysis. For example, one could comparatively analyse the first two Slovenian daily papers, the liberal Slovenski narod and the conservative Slovenec. A similar comparative analysis could be applied to Edinost and Soča, the two periodicals of Slovenes in Italy, analysing their overlap and divergences (especially considering their merger intentions). A much more demanding research could consider analysing differences in advertisements, given that they feature prominently even in keyword analysis. The task is complex because it is extremely difficult to set the boundaries of individual advertisements. The problem could be approached by treating periodicals as images (van Galen, 2023) and using neighbour search to find similar advertisements. LLMs can be used for all of the above tasks, which shows how this technology will revolutionise historical research in the future, especially when dealing with lower-quality corpora.

6 ACKNOWLEDGMENTS

I would like to sincerely thank dr. Nikola Ljubešić and Filip Dobranić for their invaluable contribution to this work. The work described in this paper was funded by the Slovenian Research Agency research programme P6-0436: Digital Humanities: resources, tools and methods (2022–2027) and the DARIAH-SI research infrastructure.

REFERENCES

Amon, S. (2008). Vloga slovenskega časopisja v združevanju in ločevanju slovenske javnosti od 1797-1945. *Javnost*, *15*, S9-S24.

Anonymous, L. (1885). Slovenski časopisi leta 1885. *Ljubljanski zvon*, *5*, 631-635.

Chow, E. H. C. (2024). *An experiment with Gemini Pro LLM for Chinese OCR and metadata extraction.* Retrieved April 5, 2024, from https://digitalorientalist.com/2024/04/05/an-experiment-with-gemini-pro-llm-for-chinese-ocr-and-metadata-extraction/

Darovec, D. (2023). *Pregled zgodovine Istre*. Koper: Založba Annales.

Dobranić, F., Evkoski, B., & Ljubešić, N. (2023). *Corpus of Slovenian periodicals (1771-1914) sPeriodika 1.0.* Slovenian language resource repository CLARIN.SI http://hdl.handle.net/11356/1881

Dobranić, F., Evkoski, B., & Ljubešić, N. (2024, May). A lightweight approach to a giga-corpus of historical periodicals: The story of a Slovenian historical newspaper collection. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 695–703). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.61

Dović, M. (2006). Literatura in mediji v Jurčičevem času. *Slavistična revija*, *54*(4), 543–557.

Dović, M. (2023). Anatomy of the “deathly silence”: Slovenian newspapers in Carniola and the pre-March censorship. *Neohelicon*, *50*(2), 543–560.

Ehrmann, M., Bunout, E., & Düring, M. (2019, September). Historical newspaper user interfaces: A review. In *85th IFLA General Conference and Assembly (IFLA).* Zenodo.

Ehrmann, M., Düring, M., Neudecker, C., & Doucet, A. (2023). Computational approaches to digitised historical newspapers (Dagstuhl seminar 22292). *Dagstuhl Reports*, *12*(7), 112–179. https://drops.dagstuhl.de/entities/document/10.4230/DagRep.12.7.112 doi: 10.4230/DagRep.12.7.112

Eiselt, I. (2015). Newspapers in the National and University Library in Slovenia–Access model. *Review of the National Center for Digitization*, *26*, 77–85.

Garcia, G. G., & Weilbach, C. (2023). If the sources could talk: Evaluating large language models for research assistance in history. In A. Šeļa, F. Jannidis, & I. Romanowska (Eds.), *Proceedings of the Computational Humanities Research Conference 2023* (p. 616-638).

Hengchen, S., Ros, R., Marjanen, J., & Tolonen, M. (2021). A data-driven approach to studying changing vocabularies in historical newspaper collections. *Digital*

Scholarship in the Humanities, *36*(Supplement 2), ii109-ii126. doi: 10.1093/llc/fqab032

Ilich, M. (1999). Nekaj o modi v slovenskem časopisju na prelomu stoletja (1895-1915). *Zgodovina za vse*, *6*, 98-108.

Jezernik, B. (2022). Katoliška duhovščina na prelomu devetnajstega in dvajsetega stoletja in proces modernizacije na Slovenskem. *Traditiones*, *51*(1), 103–145.

Kermavner, D. (1962). Drugi slovenski socialnodemokratski listi. *Kronika*, *10*, 80-89.

Kettunen, K., & Pääkkönen, T. (2016). Measuring lexical quality of a histori- cal Finnish newspaper collection – Analysis of garbled OCR data with basic language technology tools and means. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 956–961). European Language Resources Association (ELRA). https://aclanthology.org/L16-1152

Kilgarriff, A. (2009). Simple maths for keywords. In *Proc. corpus linguistics* (Vol. 6).

Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X., Liu, C., Jin, L., & Bai, X. (2024). *On the hidden mystery of OCR in large multimodal models.* https://arxiv.org/abs/2305.07895

Marjanen, J., Kurunmäki, J., Pivovarova, L., & Zosa, E. (2020). The expansion of isms, 1820-1917: Data-driven analysis of political language in digitized newspaper collections. *Journal of Data Mining & Digital Humanities*.

Marjanen, J., Vaara, V., Kanner, A., Roivainen, H., Mäkelä, E., Lahti, L., & Tolonen, M. (2019). A national public sphere? Analyzing the language, location, and form of newspapers in Finland, 1771–1917. *Journal of European Periodical Studies*, *4*(1).

Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., & Tolonen, M. (2021). Topic modelling discourse dynamics in historical newspapers. In S. Reinsone, I. Skadiņa, A. Baklāne, & J. Daugavietis (Eds.), *Digital humanities in the Nordic countries 2020* (pp. 63–77). CEUR-WS.org. http://dig-hum-nord.eu/conferences/ dhn2020/

Marušić, B. (2005). *Pregled politične zgodovine Slovencev na Goriškem: 1848-1899*. Nova Gorica: Goriški muzej.

Marušič, B. (2007). Izbor vesti o Istri v slovenskem časopisju do leta 1880. *Annales*, *17*.

Mayer, A. I. L., Gutierrez-Vasques, X., Saiso, E. P., & Salmi, H. (2022). Underlying sentiments in 1867: A study of news flows on the execution of Emperor Maximilian I of Mexico in digitized newspaper corpora. *Digital Humanities Quarterly*, *16*(4).

Mihelič, S. (1948). Kmetijska družba in ustanovitev "Novic". *Slavistična revija*, *1*(1/2). http://www.dlib.si/?URN=URN:NBN:SI:DOC-HU751MKO

Park, J., & Cordell, R. (2023). A quantitative discourse analysis of Asian workers in the US historical newspapers. In M. Hämäläinen et al. (Eds.), *Proceedings of the Joint 3rd international conference on natural language processing for digital humanities and 8th International workshop on computational linguistics for Uralic languages* (pp. 7–15). Association for Computational Linguistics. https://aclanthology.org/2023.nlp4dh-1.2

Pedrazzini, N., & McGillivray, B. (2022). Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers. In M. Hämäläinen, K. Alnajjar, N. Partanen, & J. Rueter (Eds.), *Proceedings of the 2nd international workshop on natural language processing for digital humanities* (pp. 85–95). Association for Computational Linguistics. https://aclanthology.org/2022.nlp4dh-1.12

Pivovarova, L., Zosa, E., & Marjanen, J. (2019). Word clustering for historical newspapers analysis. In C. Vertan, P. Osenova, & D. Iliev (Eds.), *Proceedings of the workshop on language technology for digital historical archives* (pp. 3–10). Varna, Bulgaria: INCOMA Ltd. https://aclanthology.org/W19-9002 doi: 10.26615/978-954-452-059-5_002

Schoots, J. (2023). Analyzing political formation through historical isiXhosa text analysis: Using frequency analysis to examine emerging African national- ism in South Africa. In R. Mabuya, D. Mthobela, M. Setaka, & M. Van Zaanen (Eds.), *Proceedings of the fourth workshop on resources for african indigenous languages (rail 2023)* (pp. 65–75). Association for Compu- tational Linguistics. https://aclanthology.org/2023.rail-1.8 doi: 10.18653/v1/ 2023.rail-1.8

Stergar, N. (1977). Narodnostno vprašanje v predmarčnih letnikih Bleiweisovih Novic. *Kronika (Ljubljana)*, *25*(3). http://www.dlib.si/?URN=URN:NBN:SI:DOC-WWIM1UTI

Steska, V., & Stelè, F. (2013). *Magolič, srečko (1860–1943).* Slovenska akademija znanosti in umetnosti, Znanstvenoraziskovalni center SAZU. http://www.slovenska-biografija.si/oseba/sbi339057/#slovenski-biografski-leksikon

Strange, C., McNamara, D., Wodak, J., & Wood, I. (2014). Mining for the meanings of a murder: the impact of OCR quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*.

Thomas, A., Gaizauskas, R., & Lu, H. (2024). Leveraging LLMs for post-OCR correction of historical newspapers. In *Proceedings of the third workshop on language technologies for historical and ancient languages (lt4hala)@ lrec-coling-2024* (pp. 116–121).

van Galen, Q. (2023). The page is an image again: Bleedmapping as an analysis technique for historical newspapers. *DHQ: Digital Humanities Quarterly*, *17*(1).

Verheul, J., Salmi, H., Riedl, M., Nivala, A., Viola, L., Keck, J., & Bell, E. (2022). Using word vector models to trace conceptual change over time and space in historical newspapers, 1840–1914. *Digital Humanities Quarterly*, *16*(2).

Zorn, T. (1987). Odmevnost jezikovnega vprašanja v listu Slovenski pravnik v letih 1871-1918. *Kronika*, *35*, 146-155.

Štepec, M. (1987). Zločin v slovenskem časopisju v 80. letih 19. stoletja. *Kronika*, *35*, 30-38.

KORPUSNO-JEZIKOSLOVNA ANALIZA KORPUSA SPERIODIKA

Prispevek predstavi računsko analizo sPeriodika, zgodovinskega korpusa sloven-skih periodičnih publikacij od leta 1771 do 1914. Avtorica se osredotoči na de-set pomembnih časopisov iz korpusa, pri čemeruporabi analizo ključnih besed, štetje pogostosti besed in konkordančno analizo za karakterizacijo vsebine in zgodovinskega razvoja slovenskega jezika. Študija opisuje značilnosti časopisov z računalniškimi metodami ter ugotovitve povezje z obdobjem intenzivnega ob-likovanja naroda po marčni revoluciji leta 1848. Poleg tega obravnava izzive, ki jih povzroča slaba kakovost optičnega prepoznavanja znakov (OCR) v zgodovin-skih dokumentih. Rezultati so trojni: 1) kvantitativni opis izbranih časopisov, 2) vpogled v zgodovinski razvoj slovenskega jezika, 3) analiza narave napak OCR v korpusu. Analiza ključnih besed razkriva specifične tematske usmer-itve časopisov, kot so kmetijstvo, pedagogika, podlistki in oglaševanje. Prav tako poudarja vlogo časopisov pri oblikovanju naroda. Študija prispeva k po-dročju digitalne humanistike s prikazom, kako lahko računalniška orodja odkri-jejo zgodovinske vpoglede iz digitaliziranih besedilnih podatkov, kljub omejitvam tehnologije OCR.

Keywords: zgodovinski časopisi, analiza ključnih besed, OCR napake, korpusno jezikoslovje