The slWaC Corpus of the Slovene Web

Tomaž Erjavec
Dept. of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, Ljubljana
E-mail: tomaz.erjavec@ijs.si

Nikola Ljubešić
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučiča 3, Zagreb, Croatia
E-mail: nikola.ljubesic@ffzg.hr

Nataša Logar
Faculty of Social Sciences, University of Ljubljana
Kardeljeva ploščad 5, Ljubljana
E-mail: natasa.logar@fdv.uni-lj.si

Keywords: language technologies, corpus linguistics, World Wide Web, Slovene language

Received: February 25, 2015

The availability of large collections of text (language corpora) is crucial for empirically supported linguistic investigations of various languages; however, such corpora are complicated and expensive to collect. In recent years corpora made from texts on the World Wide Web have become an attractive alternative to traditional corpora, as they can be made automatically, contain varied text types of contemporary language, and are quite large. The paper describes version 2 of slWaC, a Web corpus of Slovene containing 1.2 billion tokens. The corpus extends the first version of slWaC with new materials and updates the corpus compilation pipeline. The paper describes the process of corpus compilation with a focus on near-duplicate removal, presents the linguistic annotation, format and accessibility of the corpus via Web concordancers. It then investigates the content of the corpus using the method of frequency profiling, by comparing its lemma and part-of-speech annotations with three corpora: the first version of slWaC, with Gigafida, the one billion word reference corpus of Slovene, and KRES, the hundred million word reference balanced corpus of Slovene.

Povzetek: Dostopnost velikih zbirk besedil (jezikovnih korpusov) je nujna za empirično podprte jezikoslovne raziskave posameznih jezikov, vendar pa je izdelava takih korpusov draga in zamudna. Korpusi besedil, zajetih s spleta, so v zadnjem času postali popularen vir jezikovnih vsebin, saj jih lahko zgradimo avtomatsko, vsebujejo pester nabor sodobnih besedilnih zvrsti in so zelo veliki. Prispevek predstavlja drugo različico korpusa slWaC, spletnega korpusa slovenščine, ki vsebuje 1,2 milijardi pojavnic. Korpus dopolnjuje prvo različico slWaC z novimi besedili, pridobljenimi z izboljšanimi orodji za zajem. V prispevku opišemo izdelavo korpusa s poudarkom na odstranjevanju podobnih vsebin ter jezikoslovno označevanje, format korpusa in njegovo dostopnost prek konkordančnika. Nato raziščemo vsebino korpusa z uporabo metode frekvenčnega profila, pri katerem leme in oblikoskladenjske oznake druge različice korpusa slWaC primerjamo s tremi korpusi: s prvo različice korpusa slWaC, z referenčnim korpusom Gigafida, ki vsebuje milijardo besed, in s stomilijonskim referenčnim uravnoteženim korpusom KRES.

1 Introduction

Large collections of digitally stored and uniformly encoded texts – language corpora – have for a number of years been the basic data resources that linguists, including lexicographers, have used for their investigations into language and for making dictionaries. However, the traditional way of compiling corpora, which involved acquiring texts from authors and publishers, which exists in many disparate formats, was very expensive in terms of time and labour.

With the advent of the Web, a vast new source of linguistic information has emerged. The exploitation of this resource has especially gained momentum with the WaCky initiative [1], which has popularised the concept of "Web as Corpus". It has also made available tools for compiling such corpora and produced large WaC corpora for a number of major European languages. Now such corpora are also being built for the so called smaller languages, such as

Norwegian [8], Czech [18] and Serbian [11], moving the concept of a "large corpus" for smaller languages up to the 1 billion token frontier.

As Web corpus acquisition is much less controlled than that for traditional corpora, the necessity of analysing their content gains in significance. The linguistic quality of the content is mostly explored through word lists and collocates [1] while the content itself is explored using unsupervised methods, such as clustering and topic modelling [17].

For Slovene, a Web corpus has already been built [12]. However, the first version of slWaC (hereafter slWaC$_1$) was rather small, as it contained only 380 million words. Furthermore, it contained domains from the Slovene top-level domain only, i.e. only URLs ending with ".si" were harvested. In the meantime, hrWaC, the Croatian Web corpus had already moved to version 2, touching the 2 billion token mark, and Web corpora for Serbian and Bosnian were built as well [11], all of them passing the size of slWaC$_1$, making it high time to move forward also with slWaC.

This paper presents version 2 of slWaC (hereafter slWaC$_2$) which tries to overcome the limitations of slWaC$_1$: it extends it with a new crawl, which also includes well known Slovene Web domains from other top-level domains, and introduces a new pipeline for corpus collection and cleaning, resulting in a corpus of 1.2 billion tokens with removed near-duplicate documents and flagged near-duplicate paragraphs.

The rest of the paper is structured as follows: Section 2 presents the corpus construction pipeline, Section 3 introduces the linguistic annotation of the corpus, its format and its availability for on-line concordancing, Section 4 investigates the content of the corpus, by comparing it to slWaC$_1$, to the balanced corpus of Slovene KRES, and the reference corpus of Slovene Gigafida, while Section 5 gives some conclusions and directions for future work.

2 Corpus construction

2.1 Crawling

For performing the new crawl we used the SpiderLing crawler[1] with its associated tools for guessing the character encoding of a Web page, its content extraction (boilerplate removal), language identification and near-duplicate removal [19].

The SpiderLing crawler uses the notion of *yield rate* to optimize the crawling process regarding the amount of unique textual material retrieved given the overall amount of data retrieved. Yield rate is calculated for each Web domain as the ratio of bytes of text contributed to the final corpus and the bytes retrieved from that domain. Web domains with a yield rate under a predefined threshold are discarded from further crawling, thereby focusing the remaining crawl on the domains where more unique textual material is to be found. SpiderLing has two predefined yield rates that control when a low-yield-rate Web domain is blacklisted; we used the lower one which is recommended for smaller languages.

As seed URLs we used the home pages of Web domains obtained during the construction of slWaC$_1$ and additionally 30 well known Slovene Web domains, which are outside the .si top-level domain.

The crawl was run for 21 days, with 8 cores used for document processing, which includes guessing the text encoding, text extraction, language identification and physical duplicate removal, i.e. removing copies of identical pages which appear under different URLs. After the first 14 days there was a significant decrease in computational load, showing that most of the domains had been already harvested and that the process of exhaustively collecting textual data from the extended Slovene top-level domain was almost finished.

After completing the crawling process, which already included document preprocessing, we merged the new crawl with slWaC$_1$. We added the old dataset to the end of the new one, thereby giving priority to new data in the following process of near-duplicate removal. It should be noted that the corpus can, in cases when the content has changed, contain two texts with the same URL but with different crawl dates.

2.2 Near-duplicate removal

We performed near-duplicate identification both on the document and the paragraph level using the onion tool[2] with its default settings, i.e. by calculating 5-gram overlap and using the 0.5 duplicate content threshold. We removed the document-level near-duplicates entirely from the corpus, while keeping paragraph-level near-duplicates, labelling them with a binary attribute on the `<p>` element. This means that the corpus still contains the (near)duplicate paragraphs, which is advantageous for showing contiguous text from Web pages, but if, say, language modelling for statistical machine translation were to be performed [10], near-duplicate paragraphs can easily be removed.

The resulting size of the corpus (in millions of tokens) after each of the three duplicate removal stages is given in Table 1. We compare those numbers to the ones obtained on the Croatian, Bosnian and Serbian domains [11], showing that the second versions of the corpora (hrWaC and slWaC), which merge two crawls obtained with different tools and were collected three years apart, show a smaller level of reduction (around 30%) at each step of near-duplicate removal, while the first versions of corpora (bsWaC and srWaC), obtained with SpiderLing only and in one crawl, suffer more data loss in this process (around 35-40%).

[1] http://nlp.fi.muni.cz/trac/spiderling

[2] https://code.google.com/p/onion/

	PHY	DND	PND	R1	R2
slWaC$_2$	1,806	**1,258**	**895**	0.31	0.29
hrWaC$_2$	2,686	1,910	1,340	0.29	0.30
bsWaC$_1$	722	429	288	0.41	0.33
srWaC$_1$	1,554	894	557	0.42	0.37

Table 1: Sizes of the Web corpora in millions of tokens after removing physical duplicates (PHY), document near-duplicates (DND) and paragraph near-duplicates (PND), with the reduction ratio (R1 and R2) after the DND and subsequent PND steps.

2.3 Linguistic annotation

slWaC$_2$ was tagged and lemmatised with ToTaLe [4] trained on JOS corpus data [5]. However, it should be noted that ToTaLe had been slightly updated, so in particular the tokenisation of slWaC$_1$ and slWaC$_2$ at times differs. The morphosyntactic descriptions (MSDs) that the words of the corpus are annotated with follow the JOS MSD specifications, however, these do not define a tag for punctuation. As practical experience has shown this to be a problem, we have introduced a punctuation category and MSD, named "Z" in English and "U" in Slovene.

3 Overview of the corpus

3.1 Size of the corpus

Table 2 gives the size of slWaC$_2$, showing separately the amount of information from the 2011 crawl, from the 2014 crawl, and overall amount of information. For each of the counted elements we give the size of the corpus after removing document near-duplicates (DND from Table 1), and for the corpus which has also paragraph near-duplicates removed (PND).

Starting with the number of domains, it can be seen that the new crawl produced less domains than the first one, due to a large number (of the complete space of URLs) of static domains being removed in the physical deduplication stage (PHY). Nevertheless, the complete corpus has, in comparison to slWaC$_1$, about 12,000 new domains. Observing the URLs, we note that the new crawl gave somewhat less URLs than the old one, and that there is little overlap between the two, i.e. about 1%: 28,315 URLs are the same from both crawls, which means that their content has changed in the last three years (and are then in the corpus distinguished by having a different crawl date).

Regarding the number of paragraphs, we give both the numbers for DND and PND, with the reduction being very similar to the reduction on the token level already expressed in Table 1, i.e. 29%. For paragraphs, sentences, words and tokens, the complete corpus is simply the sum of the items for each of the two crawls. The most important numbers are the sizes of the complete corpus in tokens, i.e. 1.25 billion words for the DND and 900 million for PND, which makes the corpus almost as large as Gigafida [13], the largest corpus of Slovene to date.

3.2 Corpus format

The annotated corpus is stored in the so called vertical format, used by many concordancing engines. This is an XML-like format in that it has opening and closing or empty (structural) XML tags, but the tokens themselves are written one per line, with the first (tab separated) column giving the token (word or punctuation) itself, the second (in our case) its lemma (or, for punctuation, again the token), the third its MSD in English and the fourth the MSD in Slovene, as illustrated by Figure 1.

```
<text domain="www.cupradan.si"
      url="http://www.cupradan.si/"
      crawled="2014">
<gap extent="1000+"/>
<p type="text" duplicate="0">
<s>
*        *        Z          U
Izmed    izmed    Sg         Dr
vseh     ves      Pg-mpg     Zc-mmr
<g/>
,        ,        Z          U
ki       ki       Cs         Vd
boste    biti     Va-f2p-n   Gp-pdm-n
delili   deliti   Vmpp-pm    Ggnd-mm
video    video    Ncmsan     Sometn
...
```

Figure 1: Vertical format of the annotated slWaC$_2$.

The example also shows a few other features of the encoding. Each text is given its URL, the domain of this URL and the year (2011 or 2014) on which it was crawled. Boilerplate removal often deletes linguistically uninteresting texts from the start (and end) of the document, which is marked by the empty gap element, which also gives the approximate extent of the text removed. The paragraphs are marked by their type, which can be "heading" or "text", while the "duplicate" attribute tells whether the paragraph is a (near) duplicate of some other paragraph in the corpus, in which case its value is "1", and "0" otherwise. Finally, we also have the empty "glue" element g, which can be used to suppress the space between two adjacent tokens in displaying the corpus.

3.3 Availability

The corpus is mounted under the noSketchEngine concordancer [15] installed at nl.ijs.si/noske. The concordancer allows for complex searches in the corpus, from concordances taking into account various filters, to frequency lexica over regular expressions.

slWaC$_2$	2011	2014	All
Domains	25,536	22,062	37,759
URLs (DND)	1,528,352	1,295,349	2,795,386
Paragraphs (DND)	7,535,453	18,303,123	25,838,576
(PND)	6,325,075	10,329,692	16,654,767
Sentences (DND)	22,615,610	50,693,747	73,309,357
(PND)	19,001,653	31,560,289	50,561,942
Words (DND)	360,273,022	718,332,186	1,078,605,208
(PND)	301,547,669	465,780,456	767,328,125
Tokens (DND)	421,178,853	837,727,874	1,258,906,727
(PND)	352,474,874	542,912,192	895,387,066

Table 2: Size of the slWaC 2.0 corpus.

We also make the corpus available for download, but not directly, mainly due to question of personal data protection. Namely, the corpus contains most of the Slovene Web, at least in the .si domain, so it also contains a lot of personal names with accompanying text. This is not such a problem with the concordancer, as similiar results on Web-accessible personal names can be also obtained by searching through Google or the Slovene search engine Najdi.si. However, being able to analyse the complete downloaded corpus enables much more powerful information extraction methods to be used, potentially leading to abuse of personal data. This is why we make the corpus available for research only, and require a short explanation of the use it will be put to. However, we make available the metadata of the corpus, in particular the list of URLs included in it, which enables other to make their own corpus on this basis.

4 Comparative corpus analysis

This section investigates how different the slWaC$_2$ corpus is from its predecessor, slWaC$_1$ and from two other corpora of Slovene [13]: the balanced reference corpus KRES, which contains 100 million words, and the reference corpus Gigafida, which contains 1.2 billion words, mostly (77%) from printed periodicals created between 1990 and 2011. The KRES corpus was sampled from Gigafida and has roughly the following structure: 35% books, 40% periodicals and 20% Internet. To establish how different these corpora are we used the method of frequency profiling [14]. We first made a frequency lexicon of the annotation under investigation (lemma or grammatical description) for slWaC$_2$ and the corpus it was compared with, and then for each item in this lexicon computed its log-likelihood (LL). The formula takes into account the two frequencies of the element as well as the sizes of the two corpora which are being compared; the greater LL is, the more the item is specific for one of the corpora. To illustrate, we give in Table 3 the first 15 lemmas with their LL score and their frequency per million words in slWaC$_1$ and slWaC$_2$, with the larger frequency in bold.

As can be noted, most of these highest LL lemmas are more prominent in slWaC$_2$; only *"biti" (to be)* and *"odstotek" (percent)* are more frequent in slWaC$_1$. Furthermore, quite a few lemmas have frequency 0 in slWaC$_1$. This is indicative of a difference in annotation between the two corpora: as mentioned, the tokenisation module of ToTaLe had been somewhat improved lately, which is evidenced in the fact that strings, such as "m2" and "3d" were wrongly split into two tokens in slWaC$_1$ but are kept as one in slWaC$_2$. It is a characteristic of LL scores that they show such divergences, which should ideally be fixed, to arrive at uniform annotation of the resources.

Lemma	LL	slWaC$_1$pm	slWaC$_2$pm
član	30,366	0.131	**0.282**
foto	23,092	0.018	**0.081**
m2	22,826	0	**0.033**
biti	22,767	**76,984**	74,493
°	21,447	0.001	**0.036**
3d	17,738	0	**0.026**
spoštovan	11,177	0.019	**0.059**
2x	11,092	0	**0.016**
tožnik	9,909	0.008	**0.036**
odstotek	9,265	**0.515**	0.393
co2	9,090	0	**0.013**
amandma	8,992	0.007	**0.031**
hvala	8,954	0.106	**0.173**
1x	8,505	0	**0.012**
ekspr	8,373	0	**0.012**

Table 3: The first 15 lemmas with highest log-likelihood scores and their frequency per million words for the comparison of the old and new version of slWaC

4.1 Lemma comparison with slWaC$_1$

The motivation behind comparing the previous and current version of slWaC was primarily to investigate what kind of text types are better represented in the new (or old) version of the corpus. Apart from the already mentioned differences in tokenisation, slWaC$_2$ is more prominent in three types of lemmas (texts).

First, there are legal texts, characterised by lemmas such as *"člen" (article,) "odstavek" (paragraph)*, *"amandma" (amendment) "tožnik" (plaintiff)*, which come predominantly from governmental domains, e.g. for "člen" mostly from uradni-list.si (official gazette), dz-rs.si (parliament), sodisce.si (courts).

Second are texts that address the reader (or, say, parliamentary speaker) directly, such as *"spoštovan" (honoured)*, *"pozdravljen" (hello)*, *"hvala" (thank you).* For "spoštovan", the most highly ranked domains are, again, the parliament, i.e. dz-rs.si, followed by vizita.si (medical help page of commercial POP.TV), delo.si (main Slovene daily newspaper), while "pozdravljen" and "hvala" come mostly from user forums. The corpus slWaC$_2$ is thus more representative in text-rich domains whose content changes rapidly and that contain user-generated content.

Third, the list contains two interesting "lemmas" with very high LL scores. The first is "ekspr" (only 19 in slWaC$_1$ but more than 9,000 in slWaC$_2$), which is the (badly tokenised) abbreviation "ekspr." meaning "expressive". It turns out that practically the only domain that uses this abbreviation is bos.zrc-sazu.si, i.e. the portal serving the monolingual Slovene dictionary SSKJ, which was newly harvested in slWaC$_2$. Similarly, the word "ino" (less than 500 in slWaC$_1$ but more than 7,000 in slWaC$_2$) turns out to be the historical form of *"in" (and).* Practically the only domain containing this word (6,000x) is nl.ijs.si, which now hosts a large library of old Slovene books. The new slWaC thus contains some extensive new types of texts coming from previously unharvested domains or domains that have had large amounts of new content added.

Finally, it is worth mentioning that the first proper noun in slWaC$_2$ appears only at position 36 in the LL list, and is "bratušek" with almost 6,000 occurrences, referring to Alenka Bratušek, the former (2013 – 2014) PM of Slovenia.

It is also instructive to see which lemmas are now less specific against slWaC$_1$. Among function words, there is less conjunction "pa" used either as an informal version of *"in" (and)* or as an adversary conjunction *but*, and there is less of *"da" (that)*, used to introduce relative clauses. The drop in the frequency of the conjunction "pa" seems to have a link in the increase of the conjunction *"in" (and)* which now demonstrates more than 22 million occurrences. Significantly lower appearance of "da" can be explained by the fact that verbs such as *"dejati" (to say)*, *"poročati" (to report)*, *"pojasniti" (to explain)*, *"povedati" (to tell)*, *"sporočiti" (to communicate)*, and *"napovedati" (to predict)*, which are usually followed by the conjunction "da" are now much less used in slWaC$_2$. Those verbs are typical for news reporting and the drop in their usage indicates a drop in the proportion of news items in the corpus.

Most of the bottom part of the LL list, of course, consists of nouns and adjectives – and all of them again confirm that harvesting of texts for slWaC$_2$ was much less focused on news portals than for the previous one. Namely, as a previous frequency profiling of Gigafida and KRES shows [2] lemmas like *"odstotek" (percent)*; *"milijon" (milion)*, *"evro" (euro)*, *"dolar" (dollar)*, *"tolar" (former Slovene currency)*; *"predsednik" (president)*, *"premier" (prime minister)*, *"država" (state)*, *"minister" (minster)*; *"ameriški" (American)*, *"britanski" (British)*, *"hrvaški" (Croatian)*, *"nekdanji" (former)*, *"leto" (year)*, *"lani" (last year)*, and *"zdaj" (now)* all typically appear in daily newspapers (or, in our case, on news portals) reporting on interior and international affairs – and, as mentioned, we found all of them at the bottom of the LL list, indicating less news in slWaC$_2$ than in slWaC$_1$ and also the shifting of major news topics (for instance from Kosovo and Iraq).

4.2 Lemma comparison with KRES

With slWaC$_2$, as with Web corpora in general, it is an interesting question of how representative and balanced they are. The easiest approach towards an answer to this question is a comparison with "traditional" reference corpora, and such experiments have been already performed, e.g. between the British Web corpus ukWaC and BNC, the British National Corpus [1]. The comparisons have shown that while Web corpora are different from classical corpora, which contain mostly printed sources, the differences are in general not great and so they can function as modern-day reference corpora.

We made a comparison between slWaC$_2$ and KRES [13], the balanced reference corpus of Slovene with 100 million words. The comparison shows that, as with slWaC$_1$, some of the differences are due to the different linguistic analyses. As mentioned, slWaC$_2$ was processed with ToTaLe, while KRES used the Obeliks tokeniser, tagger and lemmatiser [7], and the two disagree in some lemmatisations, the most prominent being *"veliko/več" (much)*, *"mogoče/mogoč" (possible)*, *"edini/edin" (only)*, *"desni/desen" (right)*, *"levi/lev" (left)*, *"volitve/volitev" (elections)*, as well as some differences in tokenisation, e.g. "le-ta" and "d.o.o." as one token or three.

Real linguistic differences concern mostly two types of lemmas. The first are highly ranked non-content words such as *"pa, tudi, sicer, ter, naš" (but, also, otherwise, and, our)*, which most likely show the bias of slWaC$_2$ texts to antithetical and intensifying sentences, sentences with binding clause elements (adducing), and sentences which either (a) describe characteristics of the institution representing itself on the Web – "naši programi, naša spletna stran" (our programmes, our Web page); (b) establish a common communication circle [9, 6] – *"naša dežela, naši plezalci" (our country, our climbers)*, or (c) include readers into a text – *"naša duhovna rast, naša pot" (our spiritual growth, our path).* The second are content lemmas, which fall into several groups: *"spleten" (Web)*, *"podjetje" (company)*, *"tekma, ekipa" (match, team)*, *"sistem, uporabnik, aplikacija" (system, user, application)*, and *"blog" (blog)*, i.e. slWaC$_2$ has more commercial, sports, and computer related texts, and, of course, text specific to the Web (blogs).

Conversely, KRES shows more lemmas to do with legal texts, such as *"član, odstavek, zakon" (article, paragraph, law)*, so that even with slWaC$_2$ having more texts of this type than slWaC$_1$, it still has much less than KRES.

KRES also has a specific group of lemmas, thematising a person in relation to another person, e.g. *"mama, oče, mož, žena" (mother, father, husband, wife)*, and verbs characteristic for interpersonal communication – *"vprašati, nasmehniti se, prikimati, zasmejati se" (to ask, to smile, to nod, to laugh)*. All these mostly come from fiction books in KRES. Two more specific lemmas are worth mentioning: *"tolar" (former Slovene currency)* shows that KRES, unlike slWaC$_2$, contains texts dating before 2007 (the changeover year to the euro in Slovenia), while "wallander", the hero of a series of detective novels, shows that KRES – at least in this instance – has too much text from a single source, here a book series.

4.3 Lemma comparison with Gigafida

Not surprisingly, a comparison between slWaC$_2$ and the Gigafida corpus showed rather similar results to the comparison between slWaC$_2$ and KRES. The top part of the list again contains content lemmas like *"spleten" (Web)*, *"aplikacija" (application)*, *"blog"*, *"uporabnik" (user)*, *"facebook"*, *"sistem" (system)*, etc., indicating slWaC$_2$ has more computer and Web related texts. However, the interesting part is the part where the two LL lists differ. First, it is obvious the Gigafida korpus has more sport related texts than KRES, therefore lemmas like *"tekma" (game)*, *"ekipa" (team)*, *"rezultat" (result)*, *"sezona" (season)*, *"trening" (training)* and *"zmaga" (victory)* are less prominent in slWaC$_2$and in KRES. The lemma *"podjetje" (company)* has a much lower LL score now as well, showing it is thematised in Gigafida in a larger proportion of texts than in KRES. Lemmas that are specific to slWaC$_2$ when we compare it to Gigafida (and not, when we compare it to KRES) are mostly non-content words, such as conjunctions *"in, ali, če" (and, or, if)*, personal pronouns *"jaz, ti" (I, you)*, and possessive personal pronouns *"moj, tvoj, vaš" (my, your$_{sg}$, your$_{pl}$)*, which show slWaC$_2$ contains more first and second person related contents most likely coming from user generated texts.

The lowest part of the LL list shows lemmas specific to Gigafida indicating Gigafida's bias towards news reporting texts thematising internal affairs, economy, and crime: *"predsednik" (president)*, *"minister" (minister)*, *"vlada" (government)*, *"občina" (municipality)*, *"prodati" (to sell)*, *"direktor" (manager)*, *"milijon" (million)*, and *"policist" (police officer)*, cf. [2].

4.4 Grammatical comparison with KRES

Apart from lemmas, it is also interesting to compare how the distribution of morphosyntactic categories of slWaC$_2$ differs from that of KRES. To this end we calculated six LL comparison scores, for uni-, bi- and trigrams of part-of-speech (PoS) and of complete morphosyntactic descriptions (MSDs).

The unigram PoS LL scores show that slWaC$_2$ has significantly more adjectives, unknown words, conjunctions, prepositions and particles, in this order. However, it has much less punctuation and numerals, and slightly less interjections. Especially with unknown words and punctuation the differences might be, at least partially, an artefact of different annotation programs. For the others, the results show that slWaC$_2$ tends more towards informal, user generated language (typical lemma for which is also *"lp"* meaning *"lep pozdrav" (best regards)* placed at position 20 in the LL list), although this conclusion is somewhat offset by the fact that it has less interjections. However, tagging interjections is notoriously imprecise, and the difference here might also be due to different taggers used. Conversely, KRES with its numerals shows a preponderance of newspaper texts, which tend to use lots of dates, times, amounts, and sports scores.

PoS bigrams again highlight the different annotation tools used. The most prominent combination in slWaC$_2$ is a numeral followed by an abbreviation, e.g. *"90 EUR, 206 kW, 298,80 m2"* but this difference is due to the fact that in slWaC$_2$ "EUR", "kW" etc. are treated as abbreviations, whereas they are common nouns in KRES. The same reasoning applies to combinations with punctuation. However, there are also legitimate combinations in the top scoring LL PoS bigrams: slWaC$_2$ has more noun + verb, adjective + noun and verb + adjective combinations, while KRES has more numeral + numeral, numeral + noun and verb + verb combinations. Scores for PoS trigrams give little new information: apart from annotation differences, the most prominent slWaC$_2$ combination is noun + noun + verb, which are mostly name + surname + predicate, e.g. *"Oto Pestner naredil"*, while the most prominent for KRES is a sequence of three numerals.

As for MSDs, the differences in unigrams in favour of slWaC$_2$ are greatest for the three unknown word types that KRES doesn't use (Xf: foreign word, Xp: program mistake and Xt: typo), followed by general adverbs in the positive degree, coordinating conjunctions, present tense first person auxiliary verb in the plural (*"smo"*) and animate common masculine singular noun in the accusative, i.e. the object of a sentence, e.g. *"otroka"*. Conversely, KRES has much more punctuation, digits, common masculine and feminine singular nouns in the nominative (i.e. subjects) and general adverbs in comparative and superlative degrees. Bigrams show that slWaC$_2$ has many more general adjective + common noun combinations in various genders and cases, while KRES has many more combinations with digits. The space of MSD trigrams is very large, and, if we discount the combinations appearing as a result of different annotations, does not show very interesting differences.

5 Conclusion

The paper presented a new version of the Slovene Web corpus, which is almost three times larger than its initial version and is made available through a powerful and freely accessible concordancer. During the construction process we focused on the content reductions obtained through near-duplicate removal, showing that both reductions to document and paragraph level remove a similar amount of content. We also compared the content of the slWaC$_2$ corpus to three other Slovene corpora (the slWaC$_1$ corpus, the balanced reference corpus KRES and the reference corpus Gigafida) with frequency profiling on lemmas and grammatical descriptions.

This comparison showed that the new version of the corpus has significantly more legal texts and specific text types, such as a dictionary and a library of historical books and (comparatively) less news. In the lemma comparison with KRES it has less legal texts but more user generated content and more commercial, sports, political and computer related texts. The comparison with Gigafida again showed slWaC$_2$ has more computer and Web related texts, while in this case sports and commercial news were no longer slWaC$_2$ specific. A larger proportion of several personal pronouns indicated a significant difference in the extent of the user generated content between the two corpora as well. The comparison of grammatical categories also shows a bias to informal writing and against newspaper items. But maybe the most surprising (although, in retrospect, quite logical) insight of the comparison using frequency profiling is that it is a very good tool to detect even slight differences in the processing pipelines used for the compared corpora, which then lead to significant differences in the (token, lemma and MSD) vocabularies.

There are several directions that our future work could take. First, by constructing the second version of two out of four existing Web corpora of South Slavic languages, two ideas have emerged: one is to build a multilingual corpus consisting of all South Slavic languages, and the second to develop a monitor corpus which would be automatically extended with new crawls in predefined time frames. The second direction is in the annotation of the corpus, where more effort should be invested in developing a gold standard processing pipeline, which could then be used to re-annotate the Slovene corpora in a unified manner. In addition, given that the Web contains a significant portion of user generated content containing non-standard language, the annotation pipeline should be extended by introducing a standardisation (normalisation) step on wordforms, similar to our approach to modernisation of historical Slovene words [16], which would then give better lemmas and MSDs, allowing for easier exploration of Web corpora.

As to the slWaC$_2$ functioning as a modern-day reference corpus of Slovene, the analysis showed considerable differences in the three corpora. In the future we therefore intend to supplement the results of the lemma comparison with the results of the topic modelling method [3, 17, 2]. From the assembled data of both methods we will be able to estimate more precisely which texts each corpus contains and, perhaps even more importantly, which texts each corpus misses. We believe the building of the next generation reference corpus of Slovene could in this way greatly benefit from the slWaC$_2$ corpus – its contents as well as its construction methodology.

References

[1] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The WaCky wide Web: a collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

[2] Nataša Logar Berginc and Nikola Ljubešić. Gigafida in slWaC: tematska primerjava. *Slovenščina 2.0*, 1(1):78–110, 2013.

[3] Michael I. Jordan David M. Blei, Andrew Y. Ng and John Lafferty. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] Tomaž Erjavec, Camelia Ignat, Bruno Pouliquen, and Ralf Steinberger. Massive multilingual corpus compilation: Acquis Communautaire and ToTaLe. *Archives of Control Sciences*, 15(3):253–264, 2005.

[5] Tomaž Erjavec and Simon Krek. The JOS Morphosyntactically Tagged Corpus of Slovene. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

[6] Monika Kalin Golob and Nataša Logar. Prostor v poročevalskem skupnem sporočanjskem krogu. *Slavistična revija*, 62(3):363–373, 2014.

[7] Miha Grčar, Simon Krek, and Kaja Dobrovoljc. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In *Zbornik Osme konference Jezikovne tehnologije*, Ljubljana, 2012. Jožef Stefan Institute.

[8] Emiliano Guevara. NoWaC: A Large Web-based Corpus for Norwegian. In *Proceedings of the NAACL HLT 2010 Sixth Web As Corpus Workshop*, WAC-6 '10, pages 1–7, 2010.

[9] Tomo Korošec. *Stilistika slovenskega poročevalstva.* Kmečki glas, Ljubljana, 1998.

[10] Nikola Ljubešić and Antonio Toral. caWaC - a Web Corpus of Catalan and its Application to Language Modeling and Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[11] Nikola Ljubešić. {bs,hr,sr}WaC: Web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the WAC-9 Workshop*, 2014.

[12] Nikola Ljubešić and Tomaž Erjavec. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In Ivan Habernal and Václav Matousek, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, Lecture Notes in Computer Science, pages 395–402. Springer, 2011.

[13] Nataša Logar, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt, and Simon Krek. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Zbirka Sporazumevanje. Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede, Ljubljana, 2012.

[14] Paul Rayson and Roger Garside. Comparing Corpora Using Frequency Profiling. In *Proceedings of the Workshop on Comparing Corpora*, pages 1–6. Association for Computational Linguistics, 2000.

[15] Pavel Rychlỳ. Manatee/bonito – a modular corpus manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, 2007.

[16] Yves Scherrer and Tomaž Erjavec. Modernizing historical Slovene words with character-based SMT. In *BSNLP 2013 - 4th Biennial Workshop on Balto-Slavic Natural Language Processing*, Sofia, Bulgaria, 2013.

[17] Serge Sharoff. Analysing Similarities and Differences between Corpora. In *Proceedings of the Seventh Conference on Language Technologies*, pages 5–11, Ljubljana, 2010. Jožef Stefan Institute.

[18] Drahomíra Spoustová, Miroslav Spousta, and Pavel Pecina. Building a Web Corpus of Czech. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. European Language Resources Association (ELRA).

[19] Vít Suchomel and Jan Pomikálek. Efficient Web Crawling for Large Text Corpora. In Serge Sharoff Adam Kilgarriff, editor, *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43, Lyon, 2012.