Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3340–3345
Marseille, 11–16 May 2020

Gigafida 2.0: The Reference Corpus of Written Standard Slovene

Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraž Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, Kaja Dobrovoljc
Jozef Stefan Institute, Ljubljana, Slovenia
Faculty of Arts, University of Ljubljana, Slovenia
{simon.krek, tomaz.erjavec, andraz.repar, nikola.ljubesic, kaja.dobrovoljc}@ijs.si
{spela.arharholdt, jaka.cibej, apolonija.gantar, iztok.kosem}ff.uni-lj.si

Abstract
We describe a new version of the Gigafida reference corpus of Slovene. In addition to updating the corpus with new material and annotating it with better tools, the focus of the upgrade was also on its transformation from a general reference corpus, which contains all language variants including non-standard language, to the corpus of standard (written) Slovene. This decision could be implemented as new corpora dedicated specifically to non-standard language emerged recently. In the new version, the whole Gigafida corpus was deduplicated for the first time, which facilitates automatic extraction of data for the purposes of compilation of new lexicographic resources such as the collocations dictionary and the thesaurus of Slovene.

Keywords: reference corpus, corpus compilation, standard language, Text Encoding Initiative, Slovene language

1. From Gigafida to Gigafida 2.0

Gigafida, reference corpus of written Slovene, belongs to (and is comparable with) the wider family of West and South Slavic (national) corpora, with representatives such as the National Corpus of Polish (Przepiórkowski et al., 2012), the Czech National Corpus (Hnátková et al., 2014), the Slovak National Corpus (Šimková et al., 2017), the Bulgarian National Corpus (Koeva et al., 2012), and the Croatian National Corpus (Tadić et al., 2009).

Gigafida was initially compiled within the framework of the "Communication in Slovene" project (2008 – 2013), together with three derived corpora: (1) Kres, a balanced subcorpus sampled from Gigafida; (2) ccGigafida (Logar et al., 2013a) and (3) ccKRES (Logar et al., 2013b), two open-access sampled subcorpora designed to facilitate the development of language technologies for Slovene (Logar et al., 2012). Furthermore, the project delivered a user-friendly concordancer which enables language research not only for linguists, but for all interested user groups who need specialised tools for analysing corpus data (Arhar Holdt et al., 2019).

Gigafida also serves as the basic source of information for Slovene language description and the compilation of digital dictionaries and lexicons (cf. Gorjanc et al., 2018). It is thus vital to ensure that the corpus is updated and upgraded. The version of Gigafida described in this paper was developed in a four-year project between 2015 – 2018, with the upgrade focusing on segments that make the greatest possible impact in relation to the initial state while relying on existing infrastructure (concordancers, taggers, etc.) developed in the previous project.

To achieve this impact, the corpus was complemented with new, contemporary texts (Section 2) and re-annotated with state-of-the-art tools for lemmatization and morphosyntactic tagging of Slovene (Section 3). Futhremore, unlike the first edition, Gigafida 2.0 has been additionally processed to ensure that it is more suitable for research on the standard version of Slovene. The corpus was cleaned of texts in which deviations from the language norm could be identified (Section 4), and duplicate texts and texts fragments were removed from the corpus (Section 5).

As a result, two versions of the new resource were prepared: Gigafida 2.0 Proto and Gigafida 2.0. The former contains 38,364 texts or 1.8 billion tokens, an increase of 29% compared to Gigafida 1.0. This version was further processed by removing duplicate texts, and other filters were applied (see Section 6). The final version, Gigafida 2.0, contains 38,310 texts and somewhere in excess of 1.1 billion words (1,134,693,333 words). In terms of size, Gigafida 2.0 is thus comparable to Gigafida 1.0 (1,186,999,699 words). Both versions of the corpus are available through several standard concordancers intended for linguists, while Gigafida 2.0 is also accessible through a custom user-friendly interface (Section 7).

2. New Corpus Material

One of the key tasks of the project was targeted collection of new materials. Based on the analysis of the previous versions of Gigafida (Logar et al. 2012: 31-41), text collection concentrated on (1) text types that were underrepresented in previous versions and (2) on texts from selected web-based publishers that would ensure larger amounts of contemporary language in the corpus, as the last texts included in Gigafida were from 2010.

2.1 Fiction

Fiction was one of underrepresented text types in the first version of Gigafida, therefore the process of collecting new texts started with the compilation of bibliographic lists of fiction works that should be targeted for the upgrade. Among the criteria for inclusion were (a) the number of borrowings of books of this genre in libraries, (b) literary prizes, (c) the inclusion of texts in the final exams of secondary schools, (f) sales records of the books in Slovene bookstores 2010 – 2014, and finally, (e) the availability of the work in digital form. The activity of increasing the amount of fiction texts was reasonably successful since Gigafida 2.0 contains 3.5% of fiction texts (compared to 2.2% in the previous version of Gigafida), and in total, we were able to collect 310 works from 9 different publishers.

2.2 Textbooks

The next group consisted of textbooks, workbooks and similar texts targeted at primary and secondary school students. One of frequently encountered problems with this type of texts is their availability, as publishers are reluctant to grant access. As a consequence, collections such as 33 e-textbooks of 9 different school subjects, available under Creative Commons 2.5 BY-NC-SA licence (https://eucbeniki.sio.si/), and similar texts, published from the last collection of texts for the Gigafida corpus, represented a valuable resource. In addition, we were able to acquire 12 history textbooks and 3 Slovenian language textbooks from one of Slovene publishing houses.

2.3 News

To ensure that corpus also includes large quantities of material from 2010 onwards, the focus of news collection were texts published by widely read Sloveian news portals (rtvslo.si, 4ur.com, siol.net, zurnal24.si, sta.si, delo.si, dnevnik.si, vecer.si, slovenskenovice.si, svet24.si) with extensive text production. Publication time of included material ranges from May 2008 to November 2018. The texts were collected using the IJS Newsfeed service (Trampuš and Novak, 2012), which regularly crawls RSS feeds of various news portals and converts articles into formats that facilitate digital processing (Bušta et al., 2017).

3. Corpus Annotation

Linguistic annotation of the Gigafida 2.0 corpus was performed on the following levels: (1) tokenization and sentence segmentation, (2) annotation with morphosyntactic descriptions (MSDs) and (3) lemmatization.

At the time this corpus was finalized, there were two major pipelines available for the Slovenian language: the first was developed within the SSJ project (Obeliks, Grčar et al., 2012, hereafter "the SSJ pipeline"), and the one developed within the CLARIN.SI research infrastructure (reldi-tagger, Ljubešić and Erjavec, 2016, hereafter "CLARIN.SI pipeline"), each with their own strengths and weaknesses. While the SSJ pipeline includes explicit linguistic post-processing rules, the CLARIN.SI pipeline excels in overall annotation accuracy. This was the reason why a "meta-tagging" approach was chosen to make use of the strong sides of each pipeline.

For the first task of tokenization and sentence segmentation, the rule-based Obeliks4J tool was applied as it follows the segmentation rules applied both in the previous versions of the reference corpus, as well as the manually annotated training data.

The remaining two annotation layers were applied via a dedicated meta-tagger. The meta-tagger is based on the supervised machine learning paradigm, the predictive model being trained on 10,000 instances (tokens) where the outputs of the two basic pipelines differ, having the manually assigned information about what the correct MSD tag and lemma for the specific token are. Using the information about predicted MSD, extended part-of-speech (first two letters of the MSD tag), and morphosyntactic features (extracted from the MSD tag) from both taggers, the meta-tagger decides which MSD tag and lemma are more likely to be correct. During the development of the meta-tagger, experiments were performed on using surface forms and their suffixes, as well as features extracted from potential lemmas, but this information did not improve the meta-tagger's results. The developed meta-tagger, evaluated on the babushka-bench platform,[1] proved to be overall more accurate than any of the two solutions by themselves, with the measured MSD accuracy of 94.34 and the lemmatization accuracy of 98.66, a relative error reduction from the best single solution of 2% for MSD tagging and 19% for lemmatization. More importantly, by using the meta-tagger, we managed to combine the preferred linguistic restrictions of the SSJ pipeline and the higher accuracy of the CLARIN.SI pipeline.

4. Corpus Standardization

Before version 2.0, Gigafida included texts which were known or assumed to contain non-standard linguistic features, i.e. deviations from standard Slovene in terms of spelling, vocabulary, syntax, and style (e.g. online comments, board messages, fiction written in dialect or slang, and some regional newspapers). In Gigafida 2.0, this type of content was minimised to the greatest degree possible for a number of reasons. Firstly, more reliable sources have since been compiled for the analysis of non-standard Slovene, in particular the Janes Corpus of Internet Slovene (Fišer et al., 2018). Secondly, the number of non-standard texts in Gigafida 1.0 was relatively small to begin with and thus not representative for the analysis of non-standard linguistic features. Finally, corpus texts were not assigned any metadata on their degree of standardness, which made it impossible to exclude them from queries.

Non-standard texts were identified automatically using a regression model (Ljubešić et al., 2015), where paragraphs were first assigned standardness values between 1 (standard) and 3 (highly non-standard). This provided a more precise overview of the distribution of non-standardness within a text. Each text was thus assigned a vector of values, one for each paragraph.

This distribution of values for each text was then compared with the distribution of values within the entire corpus by applying the Kolmogorov-Smirnov test. All texts with a statistically significant deviation from the distribution in the entire corpus and with a mean value of linguistic non-standardness higher than the corpus average were included on a list of potentially problematic texts. This selection was then manually verified to create the final selection of 16 texts comprising a total of approximately 1,9 million tokens. The texts all included a high degree of some form of non-standard Slovene variant such as slang (e.g. the Slovene translation of Irvine Welsh's *Trainspotting*), regional language variants (e.g. *Novi Matajur*, the newspaper of the Slovene minority in Italy), and archaic expressions (e.g. two Slovene books of prayers and hymns). These texts were removed from the corpus.

[1] https://github.com/clarinsi/babushka-bench

In addition, approximately 160 million tokens were removed, the majority coming from web-crawled texts from 2010 and 2011. Most of these texts were online articles from news sites, but were crawled along with user comments. As there was no clear distinction in the text structure between articles and comments, all of the texts were removed.

5. Deduplication and Filtering

After its publication, the use of Gigafida demonstrated the need to remove duplicates, since printed media texts often contain duplicate or near-duplicate segments that skew statistical data analyses of the corpus as a whole. A typical example includes radio and TV programmes with identical content published in different sources, as well as (newspaper) publications summarising the same source. During the process of compiling Gigafida 2.0, deduplication was performed on all texts that had been previously included and then proceeding to newly collected texts.

The Gigafida 2.0 Proto corpus contains 38,364 texts or 1.8 billion tokens, an increase of 29% compared to Gigafida 1.0. Gigafida 2.0 Proto corpus was further processed by removing duplicate texts and very short texts. In comparison to version 1.0, all texts with non-standard linguistic elements were also eliminated. This resulted in the final version of the corpus, Gigafida 2.0, which contains 38,310 texts or somewhere in excess of 1.1 billion words (1,134,693,333 words). In terms of size, Gigafida 2.0 is thus comparable to Gigafida 1.0 (1,186,999,699 words).

De-duplication was performed on the level of paragraphs with the help of Onion (ONe Instance Only, Pomikalek, 2011), which allows for the setting of two parameters, -n and -t. The first defines the length of identical n-grams, which the program uses during de-duplication, whereas the second defines the quantity of identical n-grams within a paragraph (duplicate content threshold). An individual paragraph counts as a duplicate when it exceeds the threshold of the number of n-grams, as defined by the -t parameter, which had already been detected in previous paragraphs. Parameter -t may have values between 0 and 1, with value 1 signifying that the paragraph is 100% identical. The following settings were applied: -n=9, and -t=0.5, which reduced the corpus size by 24.9% (cf. Benko, 2013).

Apart from paragraphs, de-duplication was also performed on the level of entire texts. When the text contained more than 95% of duplicate paragraphs, the whole text was removed from the de-duplicated corpus. Furthermore, in another filtering step, 6,915 texts were removed from the corpus; this included texts which: 1) did not include letters the Slovene letters ‘č’, ‘ž’ and ‘š’, which eliminated texts with character encoding issues, or where the authors failed to use standard spelling (see Section 4); and 2) which contained fewer than 500 characters, as these hardly contain any text, yet all the metadata skew corpus statistics as well as lead to a greater number of corpus files.

6. Corpus Composition

As seen in Table 1 and Figure 1, Gigafida 2.0 is mainly comprised of newspapers (amost half of all words), online texts (about one quarter), and periodicals (one sixth), with smaller percentages coming from non-fiction, fiction, and other genres, with the percentages comparable to those of Gigafida 1.0. The greatest discrepancy occurs with online texts, with Gigafida 2.0 containing about 12% more. The new version has fewer newspaper texts (a decrease of 8%) and periodicals (a decrease of 5%) and a slightly larger proportion of fiction (an increase of 1.5%). However, it should be mentioned that online texts also include news: in the previous version, these were only included in the Newspapers category. In Gigafida 2.0, they were obtained in digital form through the IJS Newsfeed (see Section 2.3), and were thus categorised as online texts. The actual differences in the distribution of genres between the two versions are thus smaller than suggested by the percentages.

Text Type	Gigafida 1.0	Gigafida 2.0
Internet	15.6%	28.0%
Newspapers	55.9%	47.8%
Periodicals	21.5%	16.5%
Non-fiction	4.2%	3.8%
Fiction	2.0%	3.5%
Other	0.7%	0.3%

Table 1: Text type distribution in Gigafida 1.0 and 2.0 by number of tokens.

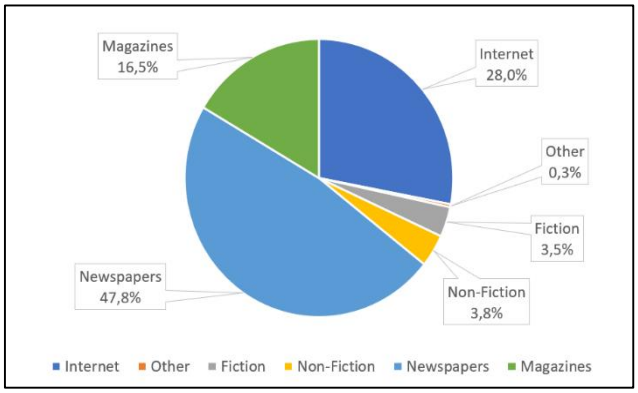

Figure 1: Text type distribution in Gigafida 2.0 by number of tokens.

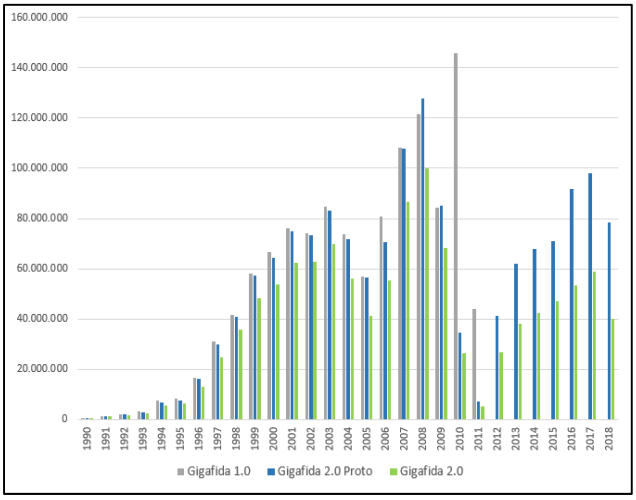

Figure 2: Distribution of millions of tokens per year in the Gigafida 1.0, Gigafida 2.0 Proto and Gigafida 2.0 corpora.

Unlike Gigafida 1.0, Gigafida 2.0 also contains texts published between the years 2012 and 2018, which represent about 27% of the Gigafida 2.0 corpus. Newspaper and periodical articles in the corpus cover the period from 1990 to 2010. Texts published after 2010 mainly include online texts; compared to the previous version, there is also a greater share of contemporary fiction.

As shown in Figure 2, the distribution of tokens in the corpus is somewhat uneven, with quite a few texts from before 1995, and a fall between 2003 – 2006 and again 2009 – 2011, which is due to the dynamics on including the texts from the Internet.

7. Corpus Concordancer

Gigafida 2.0 and Gigafida 2.0 Proto are available through various concordancers. They are included in the concordancers NoSketch Engine and KonText of the Slovene research infrastructure CLARIN.SI[2]. and the SketchEngine tool[3]. Gigafida 2.0 is also included in the CJVT portal[4] within a concordancer designed for general use.

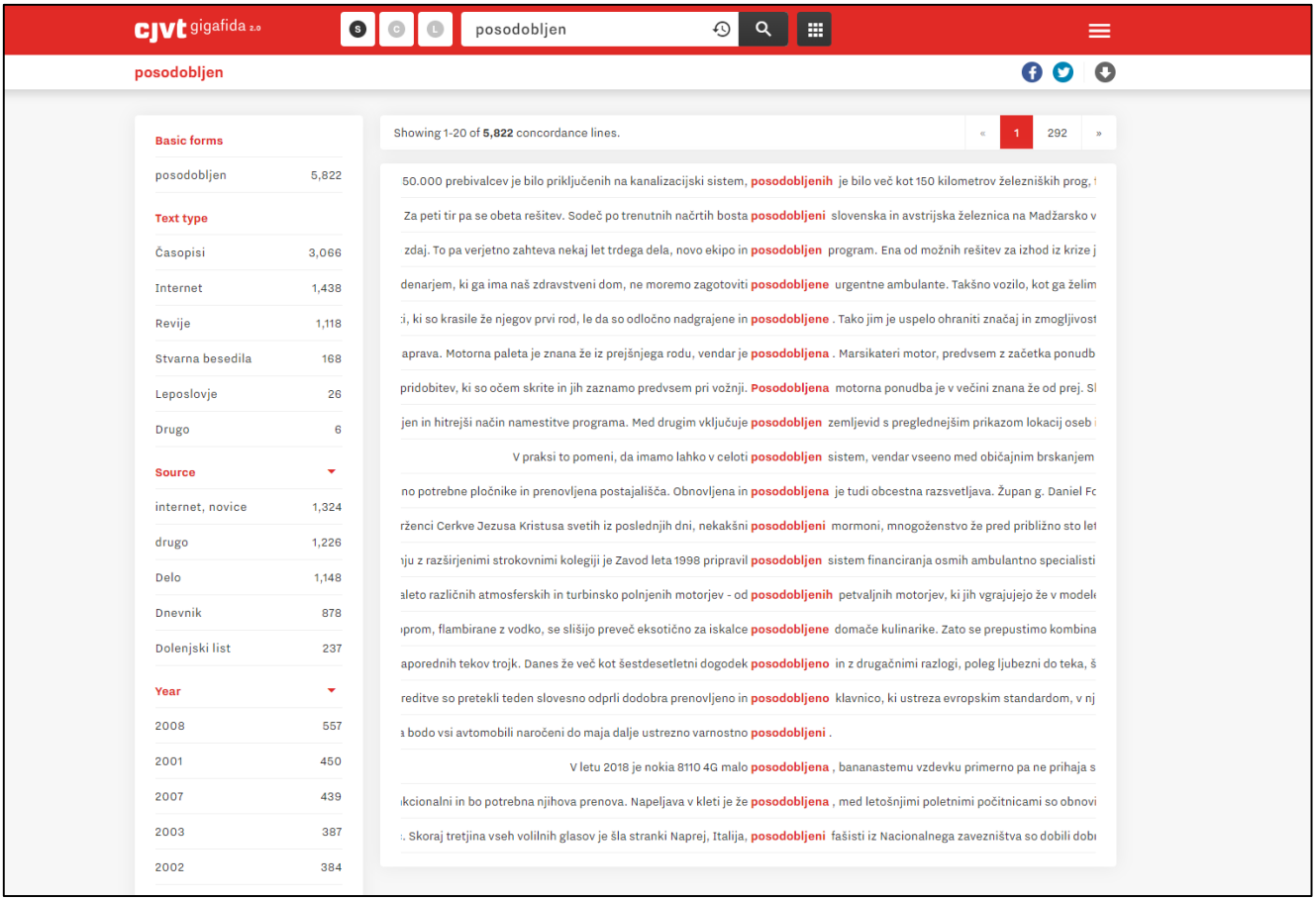

Figure 3: Concordances for posodobljen ('updated').

The concordancer for general use was launched in 2012, together with the first version of the Gigafida corpus. The program was developed with the intention to facilitate corpus use and access to all interested user groups, such as teachers, editors, translators, writers of texts for various uses etc. The design and functionality of the concordancer were evaluated a few years after the launch. Target user groups have positively rated the concordancer features, the two main highlights being the simplicity of searching for corpus data and the clean look of the interface (Arhar Holdt et al., 2019).

For Gigafida 2.0, the concordancer was upgraded. In keeping with the results of the user evaluation, all the main functionality has been retained: separate tabs for concordance and collocation searches and for generating word lists; progressive interface navigation, which is enabled by data interconnectedness (e.g. shifting from the list of collocates to concordance sequence); simple basic and advanced searches; interactive data filters; access to search history; the possibility of exporting data to other programs for further processing.

However, some interface adjustments were included (Figure 3), allowing for greater ease of search and a better overview of the results: a) buttons for shifting between different search modes are now placed conveniently close to the search window; b) shifting between tabs Search, Collocation and List initiates automatic query for the selected data type, increasing time efficiency; c) frequency data is now more accessibly categorised within filters; d) the filters allow for simultaneous choice of two or more categories as well as for a simple clearing of selected filters; e) navigation through concordance string pages allows for a direct jump to the selected page; f) instead of MI statistics,

[2] https://www.clarin.si/noske/
[3] https://www.sketchengine.eu/
[4] https://viri.cjvt.si/eng/

the tab Collocations now provides the more relevant logDice; and g) the addition of buttons for sharing content on social media. The funcionalities of the programme can be tested at https://viri.cjvt.si/gigafida/, where the concordancer is available also with English interface.

The interface was primarily developed for use on computers; however, the design was adjusted for use on tablets and smart phones, with space limitations taken into account.The inclusion of Gigafida 2.0 in the CJVT portal facilitates simple interlinking of various types of language data. By clicking a button next to the search window, the users are given the possibility to search for the selected query simply and quickly in different CJVT sources, e.g. in the Thesaurus of Modern Slovene or the Collocations Dictionary of Modern Slovene.

8. Future Work

We have presented version 2 of the Gigafida reference corpus of modern standard Slovene and the related concordancing tools for its browsing. The Gigafida 2.0 has already been used as the central language resource in the compilation of several corpus-based dictionaries of modern Slovenian, such as the Thesaurus of Modern Slovene (Arhar Holdt et al., 2018), the Collocations Dictionary of Modern Slovene (Kosem et al., 2018) and the Sloleks 2.0 Slovene Morphological Lexicon (Dobrovoljc et al., 2019). In the next step, the online browsing interfaces for these resources[5] will be updated with direct hyperlinks to the Gigafida 2.0 concordances for all relevant lemmas, word forms and collocations. This feature will enable the advanced dictionary users to perform a more detailed analysis of specific phenomena on a much larger set of authentic examples and to make use of additional functionalities provided by the Gigafida 2.0 concordancer, such as filtering by text type, source or year of publication.

The Gigafida 2.0 corpus has also been recognized as an important language resource for the development of future corpus-based grammatical descriptions of modern Slovene. As part of the ongoing national project "New grammar of contemporary standard Slovene: sources and methods", several important datasets have already emerged from the Gigafida 2.0 corpus, such as the openly available frequency lists of characters (Čibej et al. 2019a), word parts (Čibej et al. 2019c), words (Čibej et al. 2019d) and word n-grams (Čibej et al. 2019b), to enable future morphological, syntactic and lexical corpus-based descriptions of modern Slovene.

To support these endeavours in the fields of linguistics and natural language processing alike, we aim to ensure a continuous development of the Gigafida corpus in the future as well. Although this undertaking also involves continuous improvements of its content and accessibility, our immediate future work focuses on adding additional layers of grammatical annotation to the database, such as information on dependency syntactic structure, semantic roles, named entities and multi-word units, for which a large-scale manually annotated training set is already available (Krek et al., 2019).

[5] https:/viri.cjvt.si

9. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency (ARRS) within the national research programmes "Language Resources and Technologies for Slovene" (P6-0411), "Slovene Language - Basic, Contrastive, and Applied Studies" (P6-0215), "Knowledge Technologies" (P2-0103), and the national basic research project "New grammar of contemporary standard Slovene: sources and methods" (J6-8256). The project "The Upgrade of Corpora Gigafida, Kres, ccGigafida and ccKres" was financed by the Slovenian Ministry of Culture.

10. Bibliographical References

Arhar Holdt, Š., Dobrovoljc, K., and Logar, N. (2019). Simplicity matters: user evaluation of the Slovene reference corpus. *Language resources and evaluation*, 53(1): 173-190. https://doi.org/10.1007/s10579-018-9429-8

Arhar Holdt, Š, Čibej, J., Dobrovoljc, K., Gantar, P., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C., and Robnik-Šikonja, M. (2018). Thesaurus of modern slovene: By the community for the community. In Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, pages 401–410, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.

Benko, V. Data Deduplication in Slovak Corpora (2013). In Proceedings of Slovko 2013: Natural Language Processing, Corpus Linguistics, E-learning, pages 27-39.

Bušta, J., Herman, O., Jakubíček, M., Krek, S., and Novak, B. (2017). JSI Newsfeed Corpus. In Proceedings of the 9th International Corpus Linguistics Conference. Birmingham, University of Birmingham.

Fišer, D., Erjavec, T., and Ljubešić, N. (2018). The Janes project: language resources and tools for Slovene user generated content. *Language Resources and Evaluation*, 2018: 1-24. https://doi.org/10.1007/s10579-018-9425-z.

Grčar, M., Krek, S., and Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In Tomaž Erjavec and Jerneja Žganec Gros, editors, Proceedings of the 15th International Multiconference Information Society (IS 2012), pages 89-94, Ljubljana, Slovenia. Institut Jožef Stefan.

Gorjanc, V., Gantar, P., Kosem, I., Krek, S. (eds.) (2018). Dictionary of Modern Slovene: Problems and Solutions. Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.

Hnátková, M., Křen, M., Procházka, P., Skoumalová, H. (2014): The SYN-series corpora of written Czech. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 160–164. Reykjavík: ELRA.

Koeva, S., Stoyanova, I., Leseva, S., Dimitrova, T., Dekova, R., Tarpomanova, E. (2012) The Bulgarian National Corpus: Theory and Practice in Corpus Design. Journal of Language Modelling, 2012, Vol. 0, No. 1, pp. 65-110.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., and Laskowski, C. (2018). Collocations dictionary of modern Slovene. In Proceedings of the XVIII

EURALEX International Congress: Lexicography in Global Contexts, pages 989–997, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.

Ljubešić, N. and Erjavec, T. (2016). Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene. In Nicoletta Calzolari, editor, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 1527-1531, Portorož, Slovenia. European Language Resources Association.

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S., and Škrjanec, I. (2015). Predicting the level of text standardness in user-generated content. Proceedings of the 10th International Conference on Recent Advances in Natural Language Processing (RANLP 2015), pages 371–3787, Hissar, Bulgaria.

Logar, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., and Krek, S. (2012). Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba. Ljubljana: Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede.

Logar, N., Dobrovoljc, K., and Arhar Holdt, Š. (2015). Gigafida: interpretacija korpusnih podatkov. In Mojca Smolej, editor, *Slovnica in slovar - aktualni jezikovni opis (Obdobja 34).* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 467–477.

Pomikálek, J. (2011). Removing boilerplate and duplicate content from web corpora. PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech republic.

Przepiórkowski, A., Bańko, M., Górski, R. L., Lewandowska-Tomaszczyk, B., (eds.). (2012). Narodowy Korpus Języka Polskiego. Wydawnictwo Naukowe PWN, Warsaw.

Šimková, M., Gajdošová, K., Kmeťová, B., Debnár, M. (2017) Slovenský národný korpus. Texty, anotácie, vyhľadávania. Bratislava: Jazykovedný ústav Ľ. Štúra SAV – Vydavateľstvo Mikula.

Tadić, Marko (2009) New version of the Croatian National Corpus. U: Hlaváčková, Dana ; Horák, Aleš ; Osolsobě, Klara ; Rychlý, Pavel (eds.) After Half a Century of Slavonic Natural Language Processing. Masaryk University, Brno, pp. 199-205.

Trampuš, M. and Novak, E. (2012). The Internals of an Aggregated Web News Feed. Proceedings of 15th Multiconference on Information Society (IS-2012).

11. Language Resource References

Čibej, J., Arhar Holdt, Š., Dobrovoljc, K., and Krek, S. (2019a). Frequency lists of character-level n-grams from the Gigafida 2.0 corpus. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1272.

Čibej, J., Arhar Holdt, Š., Dobrovoljc, K., and Krek, S. (2019b). Frequency lists of word-level n-grams from the Gigafida 2.0 corpus. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1274.

Čibej, J., Arhar Holdt, Š., Dobrovoljc, K., and Krek, S. (2019c). Frequency lists of word parts from the Gigafida 2.0 corpus. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1275.

Čibej, J., Arhar Holdt, Š., Dobrovoljc, K., and Krek, S. (2019d). Frequency lists of words from the Gigafida 2.0 corpus. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1273.

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L., and Robnik-Šikonja, M. (2019b). Morphological lexicon Sloleks 2.0. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1230.

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, Š., Kavčič, T., Škrjanec, I., Marko, D., Jezeršek, L., and Zajc, A. (2019). Training corpus ssj500k 2.2. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1210.

Logar, N., Erjavec, T., Krek, S., Grčar, M., and Holozan, P. (2013a). Written corpus ccGigafida 1.0. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1035.

Logar, N., Erjavec, T., Krek, S., Grčar, M., and Holozan, P. (2013b). Written corpus ccKres 1.0. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1034.