Building the v4 of the Croatian National Corpus

Marko Tadić, Vanja Štefanec, Daša Farkaš

University of Zagreb Faculty of Humanities and Social Sciences
Ivana Lučića 3, 10000 Zagreb, Croatia
{marko.tadic, vstefane, dfarkas}@ffzg.unizg.hr

Abstract

It has been thirteen years since the release of the current version (v3) of the Croatian National Corpus (HNK). In terms of synchronicity in corpus linguistics, that many years may be considered quite some time. The preparatory phase for the composition of the new version of HNK (v4) has been going already for several years and in this paper we touch on several issues of concern. Apart of regular corpus parameters, e.g. text sources, text genres, coverage of language varieties, time span, we also discuss about metadata and linguistic annotation schemata. One of important technical prerequisites was the development of CorpRepo, a custom corpus data management system and file system, which enable us to do sustainable long-term maintenance of the data, and to produce newer versions of corpus more easily and more often. The selection of IPR-cleared data entails some restrictions and we give several examples of that kind of textual sources, but also discuss possible weaknesses of such approach to data selection. Regarding the linguistic annotation, the important shift is the decision to abandon the MulText East morphosyntacting descriptions and use solutions recommended by UD-initiative.

Keywords: Croatian National Corpus, corpus annotation, corpus data management

1. Introduction

It has been more than ten years since the release of the current version, version 3, of the Croatian National Corpus (HNK), in 2013 and seventeen years since the version 2.5 in 2009 (Tadić, 2009). In terms of synchronicity in corpus linguistics, that many years may be considered quite a long time since not just new texts appeared, but the distribution of types and genres as well as means and channels of text circulations in society also changed. The same goes for advances of annotation tools, since the HNK has not been re-annotated and no new annotation layers were added during all this time.

The preparatory phase for the composition of the new version of HNK (v4) has been going already for several years during which we have been considering various issues like text sources, text genres, coverage of language varieties, metadata, time span, structural and linguistic annotation schemata.

We've also been focusing on the development of the custom corpus data management system, which would enable us to do sustainable long-term maintenance of the data, and to produce newer versions more easily and more often.

In the times when sizes of large corpora exceed several dozens of billions of tokens, and LLMs are routinely used to generate textual content, it is difficult to find justification for composing a moderately-sized hand-picked national corpus. After all, corpus linguistics could be regarded as a data science in which “more the data, more the value”. We, however, argue that in fact it does make sense to embark on that endeavour because humanly curated large representative corpora should still be considered a gold standard in corpus linguistics.

The paper presents the topic of text types and related IPR issues in section 2. The section 3 is explaining the metadata approach while the section 4 describes the custom corpus data management system. In section 5 the plans for linguistic annotation are laid down and the paper ends with conclusion as section 6.

2. Text Types and IPR

It is planned that the new version of HNK will, unlike its predecessors, contain IPR-cleared texts and will be made freely available for academic purposes under a permissive license. This means that all texts will be either collected from publicly available sources or acquired from public institutions who are obliged by law to give access to textual data they are producing or managing in some way, for scientific purposes. On top of that we are in the process of negotiating the data-providing agreements with important Croatian publishers like Matica hrvatska and Lexicographic Institute Miroslav Krleža. Since they receive support from the Ministry of Culture and Media and/or Science for publishing their editions, these might be also partially available under permissive license. At this moment we can't predict the outcome of these negotiations and any projections on these text types and their size would be highly speculative. However, we will do our best to include as much fiction as possible and we already have some texts, that have been donated by the authors themselves.

This intention to use IPR-cleared data could in fact turn the expected well-balanced and

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 80–83. Palma, Mallorca. May 11, 2026.

representative corpus into a corpus that could be called an opportunistic one since it might not follow the balanced representation of different text types, genres, domains, etc. We are well aware of that possibility, but we expect that the impact and usability of freely available very large corpus can be more important than meticulously followed theoretically planned structure.

Text types that are IPR-cleared include, for example, papers from the Portal of Croatian scientific and professional journals[1] published under different versions of open access. This portal includes 572 scientific and professional journals with more than 322,000 full-text papers from all domains of science. These papers together with BA, MA and PhD theses in open access at Digital Academic Archives and Repositories[2] from different Croatian universities and from the National and University Library form the Croatian Scientific Corpus, which will be a part of the HNK v4. The estimated size of this subcorpus is more than 600 Mw.

Another example of IPR-cleared text types is the corpus of legal texts published in the official journal Narodne novine[3]. This includes texts of laws and lower-level legal documents of the Parliament, Government, regional authorities, Constitutional Court and Croatian Central Bank. Additional documents of local authorities are available in the digital form through the Central Catalogue of the Official Documents of the Republic of Croatia[4]. Texts from both sources were partially included in the Croatian MARCELL Legislative Suborpus (Váradi et al, 2020), which was 102 Mw in size at that time. However, we are planning to include also texts produced since that time until today.

Tentative list of text types and their approximate proportions for HNK v4, while the desired target size is at least 1Gw:

- newspapers and magazines 50%
- legislative and public texts 15%
- academic 15%
- literature 10%
- publicistics 5%
- mixed types 5%

3. Metadata

Given the fact that also the structure of corpus users has broadened since the current HNK v3 appeared and it nowadays includes researchers from all fields and branches of humanities as well as social sciences, special attention will be put on catering for their specific requirements. This includes a much richer document metadata description, as well as preserving and unifying general structures within a document (titles, headings, paragraphs, articles, etc.), which were rather inconsistent in HNK v3.

In the case of two mentioned sources for scientific text types (Hrčak and Dabar repositories), they use slightly different metadata schemata, so we have built a new (meta)data model in order to harmonize the description of their objects. After importing all the metadata in the database, we harvested the actual objects from the respective repositories.

Metadata describing published scientific papers contain, among others, information about the classification using the triple layered hierarchy in different scientific domains, fields and branches in accordance with the Croatian regulations on classification of domains of sciences and arts[5]. This information from metadata facilitates the generation of domain-specific subcorpora and research that will enable the comparison of texts between different scientific fields.

4. Data Management

In very large corpora data management represents a specific challenge, but we still wanted to have a sustainable data management that is flexible enough regarding the metadata control, harvesting of documents, extraction of raw text, etc. Since for corpora data management there are no universally applicable guidelines, we set our own requirements on data management at least for large systematized sources of data like institutional repositories: 1) save all metadata, 2) version-track all documents, 3) enable collaborative work, 4) provide work environment that ensures long-term sustainability of data. We developed CorpRepo, a custom database-driven software solution (web-application), which can control large number of git repositories. It enables us to: 1) ingest and parse document metadata, 2) harvest documents and store them on file-system, 3) perform necessary git repository actions (add, commit, push, pull), 4) perform automatic, semi-automatic and manual document editing, 5) calculate and re-calculate relevant statistics, 6) generate corpora based on various parameters.

4.1 Metadata ingestion and processing

Metadata is harvested through repositories' OAI-PMH interface while webapp parses metadata, extracts relevant data and stores them into database along with the full metadata record. In many cases the metadata records also contain document abstracts.

[1] https://hrcak.srce.hr/
[2] https://dabar.srce.hr/
[3] https://www.nn.hr/
[4] https://sredisnjikatalogrh.gov.hr/
[5] https://narodne-novine.nn.hr/clanci/sluzbeni/2024_01_3_69.html

4.2 Document harvesting

Webapp takes document URL (or some other permanent identifier) from the metadata record, downloads the document and saves it onto the file-system.

4.3 Text extraction

For text extraction we're using GROBID[6], an ML library for parsing and re-structuring raw documents into structured XML/TEI. Webapp takes the document from the file-system and sends it to GROBID API where text is extracted from the resulting XML/TEI and saved onto file-system, along with the original TEI

4.4 Repository management

In repository management webapp controls both, the file-system and git repositories. All changes in the file-system are version-tracked.

4.5 Data cleaning

Through the webapp user can perform semi-automatic or manual cleaning of the data. Accuracy of the text extraction process varies significantly based on the document layout and date of creation. Text extracted from PDF documents is extremely noisy, especially in Hrčak dataset.

5. Linguistic Annotation

The largest improvement will be made with the linguistic annotation. HNK v3 has been annotated only on morphosyntactic level and lacked the annotation of syntactic and semantic relations or named entities. Moreover, the annotation was performed using the Croatian MULTEXT-East tagset v4[7] (Erjavec, 2010), composed according to the specifications defined within the MULTEXT project (Dimitrova et al., 1998). With the introduction of Universal Dependencies (UD) initiative[8] (de Marneffe et al., 2021) and the inclination of the linguistic community towards the annotation standard it proposed and constantly develops, we decided to discontinue the use of MULTEXT-East tagset. We also believe that the basic UD tagset is much more legible to wider audience of users, less redundant, and much better documented. Also, given that the central tendency of UD is to provide a framework for consistent annotation of grammar across different human languages, we believe that this will increase the importance of the HNK v4 in the cross-linguistic research as well.

When it comes to the access to the corpus through a concordancing interface, we plan to maximally simplify creating even the very complex queries by creating various attributes during preprocessing. This primarily refers to annotating perifrastic verb forms, and performing traditional, non token-based lemmatization, which will enable even non experts in the field, to use the corpus in research, teaching, language-learning, or merely for information purposes.

6. Conclusion

We have presented the work-in-progress on the development of the v4 of the Croatian National Corpus (HNK v4) and several areas of interest to the main topic of this workshop. Once completed, HNK v4 will be searchable online through HR-CLARIN concordancer[9] as well as the Sketch Engine system (Kilgarriff et al., 2004). The expected date of release is the end of 2026 or beginning of 2027.

7. Acknowledgments

This research has been partially funded by the Ministry of Science, Education and Youth of the Republic of Croatia through the support to the HR-CLARIN consortium, a Croatian participation in the CLARIN ERIC.

8. Bibliographical References

de Marneffe, M.-C., Manning, C., Nivre, J. and Zeman, D. (2021). Universal Dependencies. In *Computational Linguistics* 47(2): 255--308.

Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevic and Tufiş, D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *Proceedings of COLING-ACL 1998*, pages 315--319, Montreal, Canada. ACL.

Erjavec, T. (2010). MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2544--2547, Valletta, Malta, May. European Language Resource Association (ELRA).

Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the Eleventh EURALEX International Congress*, pages 105--116, Lorient, France, July.

Ljubešić, N. and Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2019)*, pages 29--34, Florence, Italy, August.

Silić, J. (2006). *Funkcionalni stilovi hrvatskoga jezika.* Zagreb: Disput.

[6] https://github.com/kermitt2/grobid

[7] https://nl.ijs.si/ME/Vault/V4/msd/html/msd-hr.html

[8] https://universaldependencies.org/

[9] https://corpora.clarin.hr

Tadić, M. (2009). New version of the Croatian National Corpus. In Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.), *After Half a Century of Slavonic Natural Language Processing*. Brno: Masaryk University, pp. 219--228.

Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pęzik, P., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R. Krek, S., Repar, A., Rihtar, M. and Brank, J. (2020). The MARCELL Legislative Corpus. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 3761--3768, Marseille, France, May. European Language Resource Association (ELRA).

9. Language Resource References

Erjavec, Tomaž; et al., 2025, Multilingual comparable corpora of parliamentary debates ParlaMint 5.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/2004.

MARCELL Croatian Legislative Subcorpus. 2020. European Language Grid repository, https://live.european-language-grid.eu/catalogue/corpus/21358.

Tadić, Marko. 2014. Croatian National Corpus v3, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), http://hdl.handle.net/11372/LRT-233.