SciPar: A Collection of Parallel Corpora from Scientific Abstracts

Dimitris Roussis, Vassilis Papavassiliou, Prokopis Prokopidis, Stelios Piperidis, Vassilis Katsouros

Institute for Language and Speech Processing (ILSP)
Athena Research and Innovation Center, Athens, Greece
{dimitris.roussis, vpapa, prokopis, spip, vsk}@ athenarc.gr

Abstract
This paper presents SciPar, a new collection of parallel corpora created from openly available metadata of bachelor theses, master theses and doctoral dissertations hosted in institutional repositories, digital libraries of universities and national archives. We describe first how we harvested and processed metadata from 86, mainly European, repositories to extract bilingual titles and abstracts, and then how we mined high quality sentence pairs in a wide range of scientific areas and sub-disciplines. In total, the resource includes 9.17 million segment alignments in 31 language pairs and is publicly available via the ELRC-SHARE repository. The bilingual corpora in this collection could prove valuable in various applications, such as cross-lingual plagiarism detection or adapting Machine Translation systems for the translation of scientific texts and academic writing in general, especially for language pairs which include English.

Keywords: Parallel Corpus, Machine Translation, Scientific Research, European Languages

1. Introduction
The significant improvements in Machine Translation (MT) that have been achieved with the use of SMT (Statistical Machine Translation) and especially NMT (Neural Machine Translation) systems in recent years, have been mainly driven by the compilation of large-scale parallel corpora. Many of these corpora cover general language use, while domain-specific ones able to capture and model different word/term/phrase translations and styles of writing are rather scarce.

Chu and Wang (2018) highlighted that there are few or no parallel corpora for the majority of language pairs and domains, although there has been significant work in data acquisition over the last years. In fact, the legal and biomedical domains are among the few cases research has focused on to propose new methods for training domain-adapted MT systems. The broad domain of scientific research and academic texts is one of those remaining relatively under-resourced, despite its importance. This importance is becoming increasingly apparent by recent empirical evidence that the research performance of a university declines as the linguistic distance of its local language from English increases, because English is the dominant language of scientific publications (Cao et al., 2021).

Institutional repositories of universities, digital libraries and national archives contain a wealth of content and metadata from bachelor and master theses, doctoral dissertations and other scientific publications. The openly available abstracts and titles are quite often translated by their respective authors in at least 2 languages: English and their native language. To the best of our knowledge, these rich sources of parallel data have not been exploited systematically to create parallel corpora related to scientific research. The few notable exceptions concern high-resource language pairs (e.g. English-Spanish, English-Portuguese, English-French), as mentioned in Section 2.

In this paper, we report on the construction of SciPar, a multilingual parallel corpus of theses and dissertations abstracts with 9.17 million sentence pairs in 31 language pairs, covering many European languages. The data have been assembled from 86 repositories and archives (see Table 4 in Appendix) with openly available metadata. The method that we used leverages the way the metadata records are stored in those repositories, and results in well aligned sentence pairs with academic content. The fact that several repositories contain theses and dissertations from multiple universities or university departments indicates that the proposed resource is constituted by a wide variety of scientific subjects and disciplines. Thus, we believe the corpus is especially useful since other multilingual corpora from scientific articles, theses and dissertations generally -and rather unevenly- focus on the biomedical domain (see Section 2).

We think that this resource will be useful in NLP (Natural Language Processing) research and especially useful for research related to the translation of scientific texts from and to English, so as to facilitate equitable access to scientific knowledge and accelerate research.

2. Related Work
To the best of our knowledge, there are not many parallel corpora based on academic theses and dissertations. A notable dataset which is freely available in the OPUS collection (Tiedemann, 2012) is the CAPES\textsuperscript{1} parallel corpus (Soares et al., 2018), which has been extracted from the corresponding Brazilian governmental body responsible for overseeing post-graduate programs. The corpus consists of ~1.2 million sentence pairs in the English-Portuguese language pair and has been evaluated with SMT and NMT systems, as well as manual assessment.

\footnote{1 https://opus.nlpl.eu/CAPES.php}
SciELO² is a distinct -yet similar- corpus which is also available on OPUS (Tiedemann, 2012) and relies on the SciELO database of scientific articles. The corpus is based on full texts -not just abstracts- and contains aligned sentences in English, Portuguese and Spanish (Neves et al., 2016). The most recent version of the corpus contains 3.3 million sentence pairs from various scientific disciplines, includes relative metadata (journal name, subject category, etc.) of which a subset is trilingually aligned. Its quality has been evaluated automatically (with SMT systems) and manually (Soares et al., 2019). As the authors highlighted, a straightforward use case of a parallel corpus with scientific texts is cross-lingual plagiarism detection, i.e., when an author translates a text from another language and presents it as an original work.

One of the first large parallel corpora of scientific papers was the ASPEC (Asian Scientific Paper Excerpt Corpus) which consists of about 3.7 million sentence pairs in English, Japanese and Chinese collected from paper abstracts (Nakazawa, 2016). The data of ASPEC are annotated according to their scientific field (e.g. physics, biology, nuclear engineering, etc.) and have been used repeatedly in the Workshops on Asian Translation (WAT).

Most scientific-related corpora focus on the biomedical domain, as it is a recurring task of the Workshops on Machine Translation (WMT). Some examples mentioned in Névéol et al. (2018) include the EDP corpus which has been constructed from open access journals in English and French, the MEDLINE corpus from the vernacular titles of articles in English, French, Spanish and Portuguese and the BVS corpus which has been constructed from biomedical abstracts in English, Portuguese and Spanish (Soares and Krallinger, 2019).

### 3. Methodology

The process we have followed in this paper for constructing the proposed multilingual resource contains the following tasks: (a) harvesting the web pages of the institutional repositories, (b) parsing their metadata, (c) locating entries with parallel content, and (d) mining sentence pairs.

We acquired the freely available metadata from 86 institutional repositories, national archives and digital libraries³. These repositories contain entries of bachelor and master theses, PhD dissertations, as well as -to a lesser extent- other scientific publications. Most of these repositories have been built using the DSpace (Smith et al., 2003) and the EPrints⁴ open-source repository software packages.

As expected, repositories differ in the type and way they store records, even if they use the same open-source repository software. Such an example can be seen by contrasting the source code of the web pages in Figure 1 and Figure 2. In particular, a thesis abstract from DSpace@NTUA⁵ (Digital Library of the National Technical University of Athens) is displayed in Figure 1, while in Figure 2, there is a thesis abstract from HELDA⁶ (Digital Repository of the University of Helsinki).

![Figure 1: Thesis abstract from DSpace@NTUA](https://dspace.lib.ntua.gr/xmlui/)  
![Figure 2: Thesis abstract from HELDA](https://helda.helsinki.fi/)

Thus, we followed an approach which was modified for each repository accordingly. Below, we outline the basic steps of our methodology:

- Initially, we identified the pattern of each repository structure and generated the potential URLs of the item records (i.e. theses, dissertations, etc.), which we fetched as HTMLs using the GNU Wget⁷ package.
- The HTMLs of the metadata records were parsed using the Beautiful Soup⁸ Python package. Custom scripts were developed to extract the parallel titles/abstracts from them and create a document pair for each record. We used the fastText language identification software (Joulin et al., 2017) to verify the language of each document and overcome possible metadata inconsistencies. In some cases, we had to filter out documents based on a low number of tokens for both languages.
- The NLTK⁹ library (Bird et al., 2009) was used to split documents into sentences and the LASER toolkit¹⁰ (Artetxe and Schwenk, 2019) was used to mine parallel sentences from the document pairs.
- Finally, the sentence pairs for each language pair were concatenated in a single file and were filtered so as to discard alignments with limited or no use (e.g. duplicates, segments containing only digits, etc.) for training MT systems (Papavassiliou et al., 2018).

The LASER toolkit assigns a score to each sentence pair as an indicator of the strength of the alignment. In recent works, such as WikiMatrix (Schwenk et al., 2021) and

---

²https://opus.nlpl.eu/SciELO.php
³We experimented with approximately 20 additional repositories which did not yield any parallel data.
⁴https://www.eprints.org/
⁵https://dspace.lib.ntua.gr/xmlui/
⁶https://helda.helsinki.fi/
⁷https://www.gnu.org/software/wget/
⁸https://www.crummy.com/software/BeautifulSoup/
⁹https://www.nltk.org/
¹⁰https://github.com/facebookresearch/LASER
CCMatrix\textsuperscript{11} (Schwenk et al., 2020), specific values for the threshold of this score (1.04 and 1.06 respectively) were experimentally set and proposed. Considering the scope of these works, i.e., mining sentence pairs globally from large monolingual corpora, the selection of such high values seems reasonable. However, in our case we are confident that the targeted document pairs are parallel and well structured, and consequently, a lower threshold could be used. After experimentation, we found that approximately 6.3% of identified sentence pairs got a score between 0.98 and 1.04 and most of them originated from titles, which constitute very good translations. Therefore, in this work we used 0.98 as a threshold.

Unfortunately, several academic repositories do not provide metadata in a structured way (e.g. a single thesis abstract may have been written in two languages with a distinct entry for each one) or the actual textual content that we are interested in is not contained in the source code of the web pages of the repository. Additionally, even though most repositories store records with the same web address followed by a numerical ID in ascending order (e.g. \url{https://dspace.tul.cz/handle/15240/11234}), there are several repositories which do not have any underlying patterns (or at least, we could not discern any) and thus, did not yield any data. We believe that our methodology has many limitations and is clearly not compatible with all repositories; however, we also believe that the data science community could provide valuable assistance in making existing academic repositories more accessible as they constitute extremely valuable sources of high-quality data.

4. Corpus Description and Information

The resource consists of 9,172,462 sentence pairs in 31 language pairs covering 25 languages in total\textsuperscript{12}. In Table 1, we list the number of sentence pairs for each of the 24 bilingual EN-X corpora. There are significant differences in the number of parallel sentences among language pairs, partly attributed to the data availability for each language, as well as to the number and size of each country’s repositories which were used. For example, we were not able to mine EN-DA (English-Danish) sentence pairs from any of the five Danish repositories we experimented with, as we could not discern any patterns in their metadata record structure. The 4\textsuperscript{th} column of Table 1 displays the average LASER score for each language pair. The relatively high scores reflect the high quality of the generated corpora. In the 5\textsuperscript{th} and 6\textsuperscript{th} columns of Table 1, the total number of words for each language in each bilingual corpus are reported.

In Table 2, the seven bilingual corpora in language pairs not paired with English are presented. These corpora are small compared to the ones reported in Table 1 because abstracts are not usually available in a third language (besides English and the native language) and when it happens it is mainly due to regional reasons. For instance, the MK-SQ (Macedonian-Albanian) collection originated from the repository of the South East European University\textsuperscript{13} (Library “Max van der Stoel”), which include trilingual abstracts in English, Macedonian and Albanian.

The resource is related to “scientific research” and “academic writing”, as evidenced by the origins of the raw data we used. After manual inspection we found that the parallel sentences cover scientific areas as diverse as economics, social sciences, medicine, informatics, mathematics, engineering, environmental studies, gender studies, history, philosophy, etc.

Although we did not conduct an analysis to determine the direction of the translation of the abstracts in our corpus, we refer the interested reader to a survey focused on the abstract writing practices of the MEDLINE corpus conducted by Névéol et al. (2020). In this survey, it was found that authors typically write abstracts in their native language first and then translate it into English, either manually or with the aid of a machine translation method (i.e. post-editing an automatic translation). It is worth noting that the aforementioned survey concerned a biomedical corpus created from abstracts of Spanish-, Portuguese- and French-speaking authors, whereas SciPar covers a wider range of scientific areas and languages. Therefore, to say that the findings of the survey in Névéol et al. (2020) also apply to the corpus described in this paper, could only be a conjecture; albeit a plausible one.

We include examples of parallel sentences from the abstracts we harvested in Table 3. The examples demonstrate the general function and style of an abstract, i.e. to present and summarize a long study in a concise manner using a formal and descriptive writing. In scientific texts, there are usually concepts and terms with varying meanings across different sub-disciplines. For example, as we can see in the first two rows (English-Slovak parallel sentences) of Table 3, the word “beam” may take different meanings; the laser beam (lúča) in physics and the structural part of a building (nosníky) in structural engineering and architecture. Similarly, in the last two rows (English-Greek sentence pairs) of Table 3, the word “derivatives” refers to mathematical derivatives (παράγωγοι) in the 3\textsuperscript{rd} row and to financial derivatives (παρόγγοι) in the 4\textsuperscript{th} row. This feature of our resource may prove especially useful in NMT systems as they are able to learn from different contexts and handle cases as in the aforementioned examples; for example, there have been applications of multi-domain architectures which predict the domain of an input sentence with pre-trained classifiers, annotate it (e.g. with a domain tag) and translate it accordingly (Chu and Wang, 2018).

---

\textsuperscript{11} WikiMatrix and CCMatrix are both freely available on OPUS (Tiedemann, 2012).

\textsuperscript{12} List of ISO 639-1 language codes: en (English), fr (French), pt (Portuguese), el (Greek), de (German), cs (Czech), hr (Croatian), pl (Polish), sv (Swedish), es (Spanish), fi (Finnish), sl (Slovenian), lv (Latvian), lt (Lithuanian), is (Icelandic), et (Estonian), nb (Norwegian Bokmål), sk (Slovak), it (Italian), hu (Hungarian), sq (Albanian), mk (Macedonian), ru (Russian), bg (Bulgarian), nn (Norwegian Nynorsk)

\textsuperscript{13} \url{https://library.seeu.edu.mk/}

2654
Table 1: Composition of EN-X bilingual corpora

|   | L1 | L2   | Sentence Pairs | Avg. score | L1 Words | L2 Words |
|---|----|------|---------------|------------|----------|----------|
| en | bg | 2,301 | 1.148         | 47,025     | 45,272   |
| en | cs | 1,064,384 | 1.190       | 20,942,952 | 17,833,886 |
| en | de | 890,184 | 1.149       | 20,216,199 | 18,457,056 |
| en | el | 742,986 | 1.155       | 19,534,059 | 19,980,967 |
| en | es | 354,459 | 1.184       | 9,236,195  | 10,543,904 |
| en | et | 83,478  | 1.144       | 1,773,673  | 1,253,635 |
| en | fi | 457,341 | 1.120       | 9,436,244  | 6,124,499 |
| en | fr | 1,123,121 | 1.178      | 27,771,546 | 31,624,558 |
| en | hr | 806,580 | 1.194       | 19,445,123 | 16,907,339 |
| en | hu | 27,421  | 1.088       | 495,174    | 400,535   |
| en | is | 110,830 | 1.106       | 2,574,566  | 2,198,031 |
| en | it | 31,279  | 1.183       | 856,093    | 916,726   |
| en | mk | 4,940   | 1.179       | 120,131    | 117,295   |
| en | nb | 56,055  | 1.164       | 1,193,905  | 1,088,614 |
| en | nn | 2,380   | 1.196       | 44,974     | 38,363    |
| en | pl | 862,075 | 1.185       | 18,957,369 | 15,540,882 |
| en | pt | 974,167 | 1.200       | 27,278,294 | 28,790,336 |
| en | ru | 3,063   | 1.151       | 56,030     | 47,966    |
| en | sk | 60,467  | 1.213       | 1,166,108  | 965,639   |
| en | sl | 300,016 | 1.176       | 6,603,118  | 5,873,979 |
| en | sq | 7,779   | 1.196       | 212,599    | 230,419   |
| en | sv | 670,815 | 1.185       | 15,398,948 | 13,625,108 |

Table 2: Composition of bilingual corpora which do not include English

|   | L1 | L2   | Sentence Pairs | Avg. score | L1 Words | L2 Words |
|---|----|------|---------------|------------|----------|----------|
| de | es | 268  | 1.202         | 5,283      | 6,227    |
| de | fr | 281  | 1.200         | 5,345      | 6,097    |
| de | ru | 198  | 1.198         | 3,462      | 3,146    |
| es | fr | 4,915 | 1.179       | 123,929    | 121,939  |
| es | ru | 728  | 1.151         | 15,830     | 11,764   |
| fr | ru | 1,333 | 1.144       | 27,711     | 20,570   |
| mk | sq | 3,710 | 1.191        | 99,645     | 110,914  |

Besides the straightforward use of parallel corpora in MT, there are also techniques, such as data augmentation and back-translation, which leverage monolingual sentences to improve the quality of NMT systems (Sennrich et al., 2015) or better adapt them to a specific domain (Mahdieh et al., 2020), e.g. that of scientific research. The monolingual parts of our corpus may prove useful for these purposes and, in Figure 3, the total number of words for each language are presented.

Table 3: Examples of different word meanings in different scientific disciplines

Figure 3: Total words in corpus by language (horizontal axis in log-scale)
5. License and Copyright Issues
The vast majority of the texts that we have acquired and processed in order to create this resource are provided under Creative Commons (CC) licenses. Nevertheless, we should note that although the texts of some theses and dissertations are copyrighted or do not allow derivative works, the titles and abstracts by themselves constitute freely and publicly available metadata. Therefore, the resulting resource has been made publicly available via the ELRC-SHARE repository in TMX format (link to resource), under a non-standard license.

6. Conclusion and Further Research
In this work, we presented SciPar, a new parallel resource comprising 9.17M segment alignments that were generated for 31 language pairs, based on the openly available metadata on institutional repositories, digital libraries of universities and national archives. The data are related to the broad domain of scientific research, as they originated from abstracts of various fields and sub-disciplines. We believe that the corpus can prove useful in training or adapting MT systems for scientific texts, especially regarding under-resourced languages.

In the future, we aim to further augment the resource with newly published content and to exploit it in experiments involving comparison of sentence alignment algorithms, induction of bilingual lexica, and MT domain adaptation. Furthermore, we plan to categorize the sentence pairs according to scientific sub-disciplines (e.g. physics, philosophy, medicine, sociology, etc.).

7. Acknowledgements
This work has been supported by the European Language Resource Coordination (ELRC), a service contract under the EC’s Connecting Europe Facility SMART 2019/1083 programme. The access to the Okeanos cloud service provided by the Greek Research & Technology Network, is greatly appreciated.

8. Bibliographical References
Artetxe, M. and Schwenk, H. (2019). Margin-based parallel corpus mining with multilingual sentence embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3197-3203, Florence, Italy, July. Association for Computational Linguistics (ACL).
Cao, Y., Triebs, T. and Tumlinson, J. (2021). Does Language Disadvantage exist in Academia? An Empirical Analysis.
Chu, C. and Wang, R. (2018). A Survey of Domain Adaptation for Neural Machine Translation. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1304-1319.
Joulin, A., Grave, É., Bojanowski, P. and Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 427-431.
Mahdieh, M., Chen, M. X., Cao, Y. and Firat, O. (2020). Rapid Domain Adaptation for Machine Translation with Monolingual Data. arXiv preprint arXiv:2010.12652.
Nakazawa, T., Naguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S. and Isahara, H. (2016). ASPEC: Asian Scientific Paper Excerpt Corpus. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pages 2204-2208.
Névéol, A., Yepes, A. J., Neves, M. and Verspoor, K. (2018). Parallel Corpora for the Biomedical Domain. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
Névéol, A., Yepes, A. J. and Neves, M. L. (2020). MEDLINE as a parallel corpus: a survey to gain insight on French-, Spanish- and Portuguese-speaking authors’ abstract writing practice. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), pages 3676-3682.
Neves, M., Yepes, A. J. and Névéol, A. (2016). The Scielo Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pages 2942-2948.
Papavassiliou, V., Prokopidis, P. and Piperidis, S. (2018). Discovering parallel language resources for training MT engines. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
Schwenk, H., Wenzek, G., Edunov, S., Grave, E. and Joulin, A. (2019). Ccmatrix: Mining billions of high-quality parallel sentences on the web. arXiv preprint arXiv:1911.04944.
Schwenk, H., Chaudhary, V., Sun, S., Gong, H. and Guzmán, F. (2021). WikiMatrix: Mining 135m Parallel Sentences in 1620 Language Pairs from Wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Main Volume, pages 1351-1361.
Sennrich, R., Haddow, B. and Birch, A. (2015). Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709.
Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., Tansley, R. and Walker, J. H. (2003). DSpace: An open source dynamic digital repository. D-Lib Magazine, 9(1).
Soares, F., Yamashita, G. H., and Anzanello, M. J. (2018). A parallel corpus of theses and dissertations abstracts. In International Conference on Computational Processing of the Portuguese Language, pages 345-352, Springer, Cham.
Soares, F. and Krallinger, M. (2019). BVS Corpus: A Multilingual Parallel Corpus of Biomedical Scientific Texts. arXiv preprint arXiv:1905.01712.
Soares, F., Moreira, V. P. and Becker, K. (2019). A large parallel corpus of full-text scientific articles. arXiv preprint arXiv:1905.01852.
Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language and Evaluation (LREC 2012), pages 2214-2218.
## Appendix: Institutional Repositories Harvested for the Creation of the Corpus

| Name of Repository | URL | Name of Repository | URL |
|--------------------|-----|--------------------|-----|
| DSpace at Riga Stradiņš University | https://dspace.rsu.lv/ | National Archive of PhD Theses | https://dspace.esa.nl/ |
| Malmö University Electronic Publishing | http://musep.mau.se/ | Nemertes - Institutional Repository of the University of Patras | https://nemertes.library.uaf.pr/ |
| Eldorado - Repository of the TU Dortmund | https://eldorado.tu-dortmund.de/ | OpenEraTs - The institutional repository of the University of Trieste | https://www.openentras.units.it/ |
| Epsilon Archive for Student Projects - Swedish University of Agricultural Sciences | https://stud.epsilon.su.se/ | OPUS - Publication Server of the University of Stuttgart | https://elib.uni-stuttgart.de/ |
| Open access publications in the SLU publication database - Swedish University of Agricultural Sciences | https://pub.epsilon.slu.se/ | Pandemos - Digital Library, Pantheum University | http://pandemos.panteion.gr/ |
| Aalto University Publication Archive | https://aaltodoc.aalto.fi/ | Polyese - University of West Africa Institutional Repository | https://polyse Почему.рф/ |
| Amihome University of Peloponnesian Repository | https://amihome.library.aegean.gr/ | PSEPEHDA - Digital Library and Institutional Repository - University of Macedonia | https://dspace.lib.aom.gr/ |
| AMUR - Adam Mickiewicz University Repository | https://repozytorium.amu.edu.pl/ | Publication System of the University of Tubingen | https://publications.uni-tuebingen.de/ |
| APOTHESES - Institutional Repository of the Hellenic Open University | https://apottheses.eap.gr/ | PYXIDA - Institutional Repository of the Athens University of Economics and Business | http://www.pyxida.aueb.gr/ |
| CRZP - Centrul de științifică științifică pr. | https://crz.upct.ro/ | Rehnburn - Freie Universität Berlin Institutional Repository | http://refubium.fu-berlin.de/ |
| Charles University Digital Repository | https://dspace.cuni.cz/ | Repositorio institucional de la Universidad Pública de Navarra | https://dspace.cuni.cz/ |
| CRIS - Repository of the Lithuanian University of Health Sciences | https://www.luminifer.is/en/ | RepositoriUM - Institutional Repository of the University of Mihuo | http://repositori.um.shumin.u/p/ |
| Croatian Digital Dissertations Repository - National and University Library in Zagreb | https://dsci.mk/hr/ | Repository - Norwegian University of Life Sciences | https://samh.ssb.no/ |
| Czech Technical University Digital Library | https://dspace.cvut.cz/ | Repository of the Norwegian University of Science and Technology | https://mus.open.uio.no/ |
| DABAR - Digital Academic Archives and Repositories | https://dabar.wse.cz/ | Repository of the Technical University of Liberec | https://dspace.tul.cz/ |
| DEAL. pr - Research Publications (ULouvain, USL-B, UNamur) | https://deal.uclouvain.be/bp/jbore/ | Repository of the University of Liège | https://reposeurolatb.uliege.be/ |
| Digital Library - University of West Bohemia in Plzen | https://dspace.zcu.cz/ | Repository of UKDM - Sz. Cyril and Methodius University in Skopje | https://repository.ukdm.mk/ |
| Digital Library of National Technical University of Athens | https://dspace.lib.ntua.gr/ | Repository of UOS (University of Ioannina) "Omphalos" | https://dionysius.lib.uoi.gr/ |
| Digital Library of University of Maribor | https://dkum.umu.se/ | Research Information System of the University of Bambang | https://pub.epsilon.slu.se/ |
| Digital Repository of Agricultural University of Athens | http://dspace.aau.gr/ | Skemman - Digital Repository of Icelandic Universities | https://skemman.is/ |
| Digitaldræga - Library as a Service | https://www.diva-portal.org/ | ESTUDO GERAL - Digital Repository of the University of Coimbra | https://estudogeral.uc.pt/ |
| Dione - University of Piraeus | https://dione.lib.unipg.it/ | UIL Repository - Università di les Iles Balears | https://dspace.ub.cat/ |
| DSpace at University of Tartu | https://dspace.ut.ee/ | Technical University Wien Bibliothek | https://dspace.tuwien.ac.at/ |
| EMU DSpace - Estonian University of Life Sciences | https://dspace.emu.ee/ | Fens Digital Library in Tartu | https://www.fens.ee/ |
| Erdélyi Digitális Adatlab - Transylvania Digital Database | https://dspace.lib.rmu.hu/ | Theses-EN-line | https://def.archives-ouvreries.fr/ |
| eRIKA - Andrej Pysczek Medzkowski Krakow University's Repository | https://repozytorium.ka.edu.pl/ | Tepco - Open Institutional Repository of Tampere University | https://tepo.tampere.fi/ |
| ESTUDOR - Digital Repository of the University of Santiago | https://estudor.usach.cl/ | UJ Library - University of Jyväskylä | https://ujo.lib.jyu.fi/ |
| Faculty of Humanities and Social Sciences Institutional Repository - University of Zagreb | http://hufitas.flaz.unizg.hr/ | UIZ Munin - Open Research Archive - The Arctic University of Norway | https://munin.uio.no/ |
| GNOSSIS - Institutional Repository of the University of Cyprus | https://gnossis-library.ucy.ac.cy/ | Universidade de Brasilia - Institutional Repository | https://repository.unb.br/ |
| Gothenburg University Publications Electronic Archive | https://uago.uu.se/ | Universidade de Lisboa - Institutional Repository | https://repository.ulis.pt/ |
| HELDA - Digital Repository of the University of Helsinki | https://helda.helsinki.fi/ | Universidade Federal da Paraiba - Institutional Repository | https://repository.ufpb.br/ |
| HELLEANCUS - Institutional Repository of the University of the Aegean | https://hellenicus.lib.aegean.gr/ | Universidade Federal de Santa Catarina - Institutional Repository | https://repository.ufsc.br/ |
| Herva, Landgardsstiftelsen Hospital research archive | https://www.herva.lib.fi/ | Universitätssammlung des Saarlandes - Doctor Theses Archive | https://thesis.uni-saarland.de/ |
| Institutional Repository of Democritus University of Thrace | https://dspace.lib.duth.gr/ | Universitè Verne - Università Bibliothèque | https://theses.univ-lille1.fr/ |
| Institutional repository for peer reviewed articles published in open access and doctoral dissertations by Icelandic Universities | https://opinissni.is/ | Université Strasbourg - Papyrus Institutional Repository | https://papyrus.lib.amontessori.ca/ |
| Institutional Repository of the Faculty of Mechanical Engineering and Naval Architecture (FAMENA), University of Zagreb | http://repository.fib.hr/ / | University of Bologna - Institutional Doctoral Theses Repository | https://sordototoro.unibo.it/ |
| Jagellonian University Repository | https://oiaj.uj.edu.pl/ | University of Debrecen Electronic Archive | https://dspace.unideb.hu/ |
| Karolinska Institutet Open Archive | https://openarchive.ki.se/ | University of Uppsala Repository (RULU) | https://dspace.uu.se/ |
| Kinsta - Open access institutional repository of the Cyprus University of Technology | https://kinsta.cat.ac.cy/ | University of Oslo - DUO Research Archive | https://www.duo.uio.no/ |
| Landau - University of Liptadu's institutional repository | https://landau.dlapland.fi/ | University of Warwick Repository | https://dspace.warwick.ac.uk/ |
| Library "Mas van der Stelt" - South East European University | https://library.senu.edu.mk/ | UTU-Pub - Open Institutional Repository of the University of Turku | https://www.uta.pub/ |
| Medical Academic Repository of MU-Varna | https://repository.mu-varna.bg/ | International Federation of Library Associations and Institutions | http://library.ifla.org/ |

Table 4: Names and links of institutional repositories harvested for the creation of the corpus