Title	Overview of corpora of academic texts in the CLARIN infrastructure
Version	1.0
Author(s)	Darja Fišer, Jakob Lenardič
Date	21-04-2020
Status	For distribution
Distribution	BoD, NCF, UI
ID	CE-2020-1615

Table of contents

1. Introduction

In the following report, we present an overview of corpora of academic texts, focusing primarily on those that are part of the CLARIN infrastructure (i.e., they are either listed in the VLO or in the repositories of the national consortia). Corpora of academic texts contain scholarly writing, which includes research papers, essays and abstracts published in academic journals, conference proceedings, and edited volumes, theses written by students at the undergraduate and graduate levels, and scientific monographs.

The report was conducted in two steps:

(i) by manual searching the VLO and the national consortia with the following keywords: “academic corpus”, “research corpus”, “research paper corpus”, and with the keyword “thesis/theses” and narrowing the results down to corpora via faceted search.

(ii) with input provided by CLARIN UI and NC coordinators.

The full results are available in this Google Docs Spreadsheet. In total, 28 corpora of academic texts were identified, 22 of which are part of the CLARIN infrastructure. In Section 2, we first provide a comprehensive list of the corpora of academic texts that are part of the CLARIN infrastructure, dividing the corpora into monolingual and multilingual resources. We subsequently describe their identification (i.e., listed in the VLO or not), their availability (download or through a concordancer), and their metadata (language, size, research discipline and period of publication, annotation, and license). In Section 3, we provide a list of identified corpora of academic texts that are not yet part of the CLARIN infrastructure, and briefly compare them to the CLARIN-integrated corpora of academic texts. Section 4 concludes the report with a discussion of the current situations and proposals for future improvements.

2. Corpora of academic texts in the CLARIN infrastructure

2.1. The corpora of academic texts in CLARIN

There are 22 corpora of academic texts in the CLARIN infrastructure. In what follows, we group them according to language (monolingual and multilingual corpora) and describe them with the following metadata:

(i) Size
(ii) Annotation
(iii) Licence
(iv) Academic discipline and period of publication
(v) Availability (online querying and/or download)
(vi) Publication

2.1.1. Monolingual corpora

Table 1 lists 19 monolingual corpora of academic texts.

Table 1: Monolingual corpora of academic texts in the CLARIN infrastructure, sorted by language

Corpus	Language	Description
Czech Sociological Review **Size:** 3 million words **Licence:** MIT	Czech	This corpus contains research papers in sociology published between 1993 and 2016. The corpus data are in the TSV format. The corpus is available for download from the LINDAT repository.
ACL Anthology Reference Corpus **Size:** 75 million tokens **Annotation:** PoS-tagged, lemmatised, author/text metadata **Licence:** CC BY-SA	English	This corpus contains research papers in computational linguistics published between 1979 and 2015. The corpus data are in the XML formal. The corpus is available for online querying through the Sketch Engine and for download from a dedicated website. For a related publication, see Bird et al. (2008).

English Scientific Text Corpus **Size:** 35 million tokens **Annotation:** PoS-tagged, lemmatised, author/text metadata, document structure **Licence:** restricted	English	This corpus contains journal articles in the following disciplines: • computer science, • computational linguistics, • informatics, • digital construction, • microelectronics, • linguistics, • biology, • mechanical engineering, and • electrical engineering. The articles were published in the 1970s, 1980s and the 200s. The corpus is available for online querying through CQPWeb (CLARIN-D distribution). For a related publication, see Degaetano-Ortlieb et al. (2013).
GENIA corpus **Size:** 437,000 words **Annotation:** PoS-tagged, syntactically parsed, annotated for terms, events, semantic relations and coreference; text metadata **Licence:** free but unspecified	English	This corpus contains journal paper abstracts in biomedicine. The corpus data are in various formats, e.g., PTB. The corpus is available for download from PORTULAN. For a related publication, see Su et al. (2008).
UH's English E-thesis corpus **Size:** 200 million tokens **Licence:** CC BY	English	This corpus contains MA and PhD theses published between 1999 and 2016. The corpus is available for online querying through the concordancer Korp (FIN-CLARIN distribution).
The Royal Society Corpus **Size:** 32 million tokens **Annotation:** PoS-tagged, lemmatised, normalised, author and document metadata **Licence:** CC BY	English (late and early modern)	This corpus contains journal articles published in Philosophical Transactions of the Royal Society of London between 1665 and 1869. The corpus is available for online querying through CQPweb and for download from the CLARIN-D repository of the University of Saarland.

		For a related publication, see Kermes et al. (2016).
Corpus of Estonian scientific texts **Size:** 5 million words **Licence:** CLARIN_ACA-NC	Estonian	This corpus contains scientific articles and PhD theses. The corpus data are in the TEI P5 format. The corpus is supposed to be available for download from META-SHARE (CELR distribution), but the download link is not working (25 February 2020).
UH's Finnish E-thesis corpus **Size:** 12.5 million tokens **Annotation:** PoS-tagged, lemmatised **Licence:** CC BY	Finnish	This corpus contains MA and PhD theses published between 1999 and 2016. The corpus is available for online querying through the concordancer Korp (FIN-CLARIN distribution).
Chambers-Le Baron Corpus of Research Articles **Size:** 1 million words **Licence:** Oxford Text Archive licence (academic use)	French	This corpus contains research papers in the following disciplines: • media/culture, • literature, • linguistics and language learning, • social anthropology, • law, • economics, • sociology and social sciences, • philosophy, • history, and • communication. The research papers were published between 1998 and 2006. This is a plain text corpus. The corpus is available for download from the Oxford Text Archive.
UH's French E-thesis corpus **Size:** 580,000 tokens **Licence:** CC BY	French	This corpus contains MA and PhD theses published between 1999 and 2016. The corpus is available for online querying through the concordancer Korp (FIN-CLARIN distribution).
UH's German E-thesis corpus **Size:** 560,000 tokens **Licence:** CC BY	German	This corpus contains MA and PhD theses published between 1999 and 2016. The corpus is available for online querying through the concordancer Korp (FIN-CLARIN distribution).

Modern Greek Dialects: scientific papers **Size:** 113,000 words **Licence:** CC BY-SA	Greek	This corpus contains scientific texts in linguistics and dialectology. This is a plain text corpus. The corpus is available for download from the CLARIN:EL repository.
OROSSIMO Corpus **Size:** 2.5 million tokens **Annotation:** marked for term candidates, "mixed structural annotation" **Licence:** CC-BY	Greek	This corpus contains academic texts in the following disciplines: • social sciences, • computer science, • economics, • linguistics, • photography, • law, • engineering, • history, • astronomy, • earth sciences and geology, • medicine and health, and • biology. The corpus is encoded in XML (XCES). The corpus is available for download from the CLARIN:EL repository. For a related publication, see Mantzari et al. (1999).
The Language of Literature and the Language of Translation (collected scientific papers) **Size:** 48,300 words **Licence:** CC BY-SA	Greek	This corpus contains journal articles in literary and translation studies. This is a plain text corpus. The corpus is available for download from the CLARIN:EL repository.
UH's Russian E-thesis corpus **Size:** 1.1 million words **Licence:** CC BY	Russian	This corpus contains MA and PhD theses published between 1999 and 2016. The corpus is available for online querying through the concordancer Korp (FIN-CLARIN distribution).
Corpus of Academic Slovene KAS 1.0	Slovenian	This corpus contains BA, MA, and PhD theses in humanities, social sciences, and natural

Size: 1.7 billion tokens **Annotation:** MSD-tagged, lemmatised, marked for bilingual and monolingual term candidates **Licence:** CLARIN.SI Licence ACA ID-BY-NC-INF-NORED 1.0		sciences published between 2000 and 2018. The corpus data are in the TEI format. The corpus is available for download from CLARIN.SI and for online querying through noSketch Engine and KonText (CLARIN.SI distribution). For a related publication, see Erjavec et al. (forthcoming).
UH's Spanish E-thesis corpus **Size:** 2.3 million tokens **Licence:** CC BY	Spanish	This corpus contains MA and PhD theses published between 1999 and 2016. The corpus is available for online querying through the concordancer Korp (FIN-CLARIN distribution).
Academic texts - humanities **Size:** 14.5 million tokens **Licence:** CC BY	Swedish	This corpus contains academic texts from humanities disciplines published between 1997 and 2012. The corpus data are in the XML format and plain text. The corpus is available for download from the SWECLARIN repository and for online querying through the concordancer Korp (SWECLARIN distribution).
Academic texts - social science **Size:** 10.8 million tokens **Annotation:** sentence segmentation **Licence:** CC BY	Swedish	This corpus contains academic texts from social sciences disciplines published between 1997 and 2012. The corpus data are in the XML format and plain text. The corpus is available for download from the SWECLARIN repository and for online querying through the concordancer Korp (SWECLARIN distribution).
UH's Swedish E-thesis corpus **Size:** 105 million tokens **Licence:** CC BY	Swedish	This corpus contains MA and PhD theses published between 1999 and 2016. The corpus is available for online querying through the concordancer Korp (FIN-CLARIN distribution).

2.1.2. Multilingual corpora

Table 2 lists 2 multilingual corpora of academic texts.

Table 2: Multilingual corpora of academic texts in the CLARIN infrastructure, sorted by language

Corpus	Language	Description

Czech and English abstracts of ÚFAL papers **Size:** 2 million words **Annotation:** document aligned **Licence:** CC BY	Czech, English	This parallel corpus contains research paper abstracts in formal and applied linguistics. For each publication, the authors were obliged to provide both the original abstract in Czech or English, and its translation into English or Czech, respectively. The corpus data are in the TSV format. The corpus is available for download from the LINDAT repository.
The KIAP corpus **Size:** 3.9 million tokens **Annotation:** PoS-tagged **Licence:** CC-BY 4.0	English, French, Norwegian	This comparable corpus contains research articles in economics, linguistics, and medicine published between 1992 and 2003. The corpus is available for online browsing through the concordancer Corpuscle (CLARINO distribution).

2.2. The overview

2.2.1. Identification

All except the following 3 (14%) corpora can be found in the VLO:

(1) OROSSIMO Corpus
(2) Modern Greek Dialects: scientific papers
(3) The Language of Literature and the Language of Translation (collected scientific papers)

All 3 corpora are listed in the CLARIN:EL repository.

2.2.2. Availability

2.2.2.1. For download and online querying

The following 5 (23%) corpora are available for download and for online querying. In the parentheses, we specify the repository from which the corpus can be downloaded and the concordancer.

(1) ACL Anthology Reference Corpus (LDC, Sketch Engine)
(2) Academic texts - humanities (SWE-CLARIN, Korp)
(3) Academic texts - social science (SWE-CLARIN, Korp)
(4) Corpus of Academic Slovene KAS 1.0 (CLARIN.SI, noSketch Engine and KonText)
(5) The Royal Society Corpus (CLARIN-D; CQPWeb)

2.2.2.2. For online querying

The following 9 (41%) corpora are available only for online querying. In the parentheses, we specify the repository with which the corpus is associated and the concordancer.

(1) UH's German E-thesis corpus (FIN-CLARIN, Korp)
(2) UH's English E-thesis corpus (FIN-CLARIN, Korp)

(3) English Scientific Text Corpus (CLARIN-D, CQPWeb)
(4) UH's Spanish E-thesis corpus (FIN-CLARIN, Korp)
(5) UH's Finnish E-thesis corpus (FIN-CLARIN, Korp)
(6) UH's French E-thesis corpus (FIN-CLARIN, Korp)
(7) UH's Russian E-thesis corpus (FIN-CLARIN, Korp)
(8) UH's Swedish E-thesis corpus (FIN-CLARIN, Korp)
(9) The KIAP corpus (CLARINO, Corpuscle)

2.2.2.3. For download

The following 7 (32%) corpora are available only for download. In the parentheses, we specify the repository from which the corpus can be downloaded.

(1) Czech Sociological Review (LINDAT)
(2) GENIA corpus (PORTULAN; however, the corpus can only be downloaded from the project webpage)
(3) Chambers-Le Baron Corpus of Research Articles (CLARIN-UK)
(4) Czech and English abstracts of ÚFAL papers (LINDAT)
(5) OROSSIMO Corpus (CLARIN:EL)
(6) Modern Greek Dialects: scientific papers (CLARIN:EL)
(7) The Language of Literature and the Language of Translation (collected scientific papers) (CLARIN:EL)

2.2.2.4. Unavailable

The following corpus is not available. In the parentheses, we specify the repository in which the corpus is listed.

(1) Corpus of Estonian scientific texts (CELR)

The corpus yields a “no download available” message despite providing a download link and an option to accept the given licence.

2.2.3. Metadata

2.2.3.1. Language

20 (91%) corpora are monolingual, containing academic texts in the following 11 languages:

(1) Czech (1)
(2) English (5)
(3) Estonian (1)
(4) Finnish (1)
(5) French (2)
(6) German (1)
(7) Greek (3)
(8) Russian (1)
(9) Slovenian (1)
(10) Spanish (1)
(11) Swedish (3)

The remaining 2 corpora are multilingual and contain academic texts in the following 2 language combinations

(1) Czech, English (parallel)
(2) English, French, Norwegian (comparative)

2.2.3.2. Size

All corpora contain information on size. All corpora report size in words or tokens.

The size of the corpora can be summarised as follows:

- 6 very small corpora (<1 million words/tokens)
- 8 small corpora (1–10 million words/tokens)
- 5 medium-sized corpora (>10-100 million words/tokens)
- 3 large corpora (>100 million words/tokens)

The largest corpus is Corpus of Academic Slovene KAS 1.0 with 1.7 billion tokens, while the smallest corpus is The Language of Literature and the Language of Translation (collected scientific papers) with 48,300 words.

2.2.3.3. Discipline and period

14 (64%) out of 22 corpora of academic texts specify the discipline of the academic texts. 10 corpora contain academic texts from multiple disciplines, while 4 corpora contain texts from a single discipline. Across all corpora, these are the most common disciplines:

(1) Linguistics (7 corpora)
(2) Computer science (3 corpora)
(3) Economics (3 corpora)
(4) Medicine (3 corpora)

16 (73%) out of 22 corpora specify the date of publication of the academic texts. The following corpora lack such information:

(1) GENIA corpus
(2) Corpus of Estonian scientific texts
(3) Czech and English abstracts of ÚFAL papers
(4) OROSSIMO Corpus
(5) Modern Greek Dialects: scientific papers
(6) The Language of Literature and the Language of Translation (collected scientific papers)

The most common publication period is 1999–2016 (all the 7 FIN-CLARIN e-thesis corpora, e.g., UH's Spanish E-thesis corpus), with only two corpora containing academic texts published before 1990:

(1) ACL Anthology Reference Corpus (1979-2015)
(2) English Scientific Text Corpus (1970s, 1980s, and 2010s)

2.2.3.4. Annotation

10 (45%) corpora provide information on annotation. It is unclear how the following 12 corpora are annotated:

(1) Czech Sociological Review
(2) UH's German E-thesis corpus
(3) UH's English E-thesis corpus
(4) UH's Spanish E-thesis corpus
(5) Corpus of Estonian scientific texts
(6) Chambers-Le Baron Corpus of Research Articles
(7) UH's French E-thesis corpus
(8) UH's Russian E-thesis corpus
(9) UH's Swedish E-thesis corpus
(10) Academic texts - humanities
(11) Modern Greek Dialects: scientific papers
(12) The Language of Literature and the Language of Translation (collected scientific papers)

In the corpora with clear metadata about annotation, these are the most common mark-up layers:

(1) PoS/MSD-tagging (7 corpora)
(2) Lemmatisation (5 corpora)
(3) Mark-up of term candidates (3 corpora)

2.2.3.5. Licence

All corpora contain information on licence.

This is the overview of the licences of the corpora:

(1) CC BY (15 corpora)
(2) CLARIN ACA (2 corpora)
(3) MIT (1 corpus)
(4) OTA licence (1 corpus)
(5) CLARIN RES (1 corpus)

3. Corpora of academic texts outside the CLARIN infrastructure

Table 3 lists 7 corpora that are not part of the CLARIN infrastructure (i.e., they are not included in a CLARIN-certified repository or listed in the VLO).

Table 3: Corpora of academic texts outside the CLARIN infrastructure, sorted by language

Corpus	Language	Description
Academic Corpus PUCV-2006 **Size:** 59 million words **Annotation:** PoS-tagged	Spanish	This corpus contains academic texts extracted from dictionaries, didactic guidelines, disciplinary texts, lectures, regulations, reports, research articles, tests, and textbooks in the following disciplines: psychology, social work, construction engineering, industrial chemistry. The corpus is not available. For a related publication, see Parodi (2010).
Academic Corpus **Size:** 3.5 million words	English	This corpus contains journal articles, book chapters, course workbooks, laboratory manuals, and course notes from the following disciplines: arts, commerce, law, and biology. This corpus is not available.
Reading Academic Text corpus **Licence:** restricted	English	This corpus contains PhD theses from the following disciplines: agriculture, psychology, food science, technology, meteorology, and history. The data are encoded in ASCII and HTML. The corpus is not available because it is restricted at present to staff and researchers at the University of Reading, and it is only available 'on-site'. However, it is possible for people outside the University to make use of the corpus on a Research Attachment arrangement.
MuchMore Springer Bilingual Corpus **Size:** 1 million tokens **Annotation:** PoS/MSD-tagged, phrase chunking, semantic class and relations, document structure **Licence:** free but unspecified	English, German	This paper contains journal paper abstracts from medical disciplines. The corpus is encoded in MuchMore XML. The corpus is available for download from a dedicated website.

Scientext corpus **Size:** 20 million words **Licence:** CC BY	French, English	This corpus contains scientific texts and argumentative essays in humanities, experimental sciences, and applied/technical sciences. The corpus is available for online querying through a dedicated webpage.
Corpus of academic Lithuanian **Size:** 9 million words **Annotation:** no linguistic annotation	Lithuanian	This corpus contains textbooks, scientific monographs, journal articles, abstracts, forewords, research reports, and master’s and PhD theses from the following disciplines: • humanities (architecture, fine art studies, ethnology, folklore studies, philosophy, linguistics, literary theory, librarianship, history, theology), • social sciences (law, political science, economics, psychology, education, management), • physical sciences (mathematics, astronomy, physics, chemistry, geography, geology and mineralogy, informatics), • biomedical sciences (medicine, dental surgery, biology, botany, agronomy, animal husbandry, pharmacy, veterinary science, forestry studies), and • technological sciences (energy studies, chemical technology, materials science, mechanics, metrology, building construction, transport technology, agricultural and environmental sciences, management and informatics). The materials were published between 1999 and 2009. The corpus is encoded in TEI 5.

		The corpus is available for online querying through a dedicated website. For a related publication, see Usonienė and Linkevičienė (2009).
Spanish-English Research Article Corpus **Size:** 5.7 million words	Spanish, English	This corpus contains journal articles published between 2000 and 2010. The corpus is unavailable.

The non-CLARIN corpora of academic texts listed in Table 3 most significantly differ from the CLARIN corpora in Tables 1 and 2 in the following two respects:

- The non-CLARIN corpora are generally unavailable for download or online searching (4 out of 7 corpora, 57%), while a significantly smaller proportion of the CLARIN corpora are unavailable (1 out of 22, 5%).
- Fewer non-CLARIN corpora (2 out of 7, 29%) contain information on licence than CLARIN corpora (100%).

4. Conclusion

In this report, we gave an overview of 22 corpora of academic texts in the CLARIN infrastructure. We presented their identification (i.e., whether they have VLO entries) and their availability (for download, online querying or both), as well as 5 types of metadata – language, size, academic discipline and publication period, annotation, and licence.

In terms of identification, 3 (14%) out of the 22 identified corpora are not listed in the VLO, all of which belong to the CLARIN:EL repository. In terms of availability, 1 (4%) out of the 22 corpora is unavailable for download and online even though the repository presents a download option. Otherwise, availability is as follows: 5 (23%) corpora are available both for online querying and download, 7 (32%) corpora are available only for download, and 9 (41%) corpora are available only for online querying. It is noteworthy that out of the 14 corpora that can be queried online, the majority (9 or 64%) are available through the concordancer Korp (FIN-CLARIN or SWE-CLARIN distribution), while the remaining 4 are available through CQPWeb (CLARIN-D distribution), Sketch Engine (LDC), noSketch Engine and KonText (CLARIN.SI distribution), and Corpuscle (CLARINO distribution).

In terms of language, most (20 or 91%) corpora are monolingual. All corpora cover languages spoken in Europe, with English being the most represented language (5 or 25% of the 20 corpora).

Information on size is available for all the 22 corpora. Over half of the corpora (14 or 64%) contain fewer than 10 million tokens/words, while 3 (14%) corpora contain more than 100 million tokens/words.

Information on annotation fares slightly worse, as it is included for 10 corpora, so fewer than half of the 22 corpora. Otherwise, the most common annotation layers are PoS/MSD-tagging (7 corpora), lemmatisation (5 corpora), and mark-up of term candidates (3 corpora). Slightly more than half of the corpora (14 or 64%) make explicit the discipline of the texts, the most common being linguistics (6 corpora). 16 (73%) corpora explicitly state the publication period of the texts, the most common being the period between 1999 and 2016. Only 3 corpora contain texts published before 1990.

Information on licence is readily included and is available for all corpora. 15 (68%) corpora are available under CC-BY, 2 (9%) under CLARIN ACA, and the rest under the MIT, OTA, and CLARIN RES licences.

Finally, we briefly compared the CLARIN corpora to 7 non-CLARIN corpora that were identified in the survey, concluding that the non-CLARIN corpora of academic texts are generally unavailable and more infrequently provide information on licence.

For future work, we will try to resolve the issues related to unavailability and work towards a more detailed description of the metadata on size, annotation, publication period and licence, filling in the gaps that were presented in the report. We will also attempt to increase the findability of the corpora in the VLO and invite the authors/curators of the non-CLARIN corpora presented in Section 3 to deposit them in certified CLARIN repositories.

5. References

Bird, Steven, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, edited by Nicoletta Calzolari, 1755–1759. http://www.lrec-conf.org/proceedings/lrec2008/pdf/445_paper.pdf.

Degaetano-Ortilieb, Stefania, Hannah Kermes, Ekaterina Lapshinova-Koltunski, and Elke Teich. 2013. SciTex – A Diachronic Corpus for Analyzing the Development of Scientific Registers. In *New Method in Historical Corpus Linguistics*, edited by Paul Bennett et al. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.686.4567&rep=rep1&type=pdf.

Erjavec, Tomaž, Darja Fišer, and Nikola Ljubešić. Forthcoming. The KAS Corpus of Slovenian Academic Writing. Submitted to *Language Resources and Evaluation*.

Kermes, Hannah, Stefania Degaetano, Ashraf Khamis, Jörg Knappen, and Elke teich. The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of LREC 2016*, edited by Nicoletta Calzolari. http://www.lrec-conf.org/proceedings/lrec2016/summaries/792.html.

Mantazi, Elena, Maria Gavrilidou, Penny Labropoulou, and George Carayannis. 1999. Collection of digital terminological resources: methodology and results. In *Proceedings of the 2nd Conference on Greek Language and Terminology*. http://www.ilsp.gr/en/research/publications?view=publication&task=show&id=328.

Parodi, Giovanni. 2010. Academic and Professional genre variation across four disciplines: exploring the PUCB-2006 corpus of written Spanish. *Linguagem em (Dis) curso*, 10 (3): 535–567. http://dx.doi.org/10.1590/S1518-76322010000300006.

Su, Jian, Xiaofeng Yang, Huaqing Hong, Yuka Tateisi, and Jun'ichi Tsujii. 2008. Coreference resolution in biomedical texts: a machine learning approach. In *Ontologies and Text Mining for Life Sciences: Current Status and Future Perspectives*, edited by Michael Ashburner, Ulf Leser, and Dietrich Rebholz-Schuhmann. http://drops.dagstuhl.de/opus/volltexte/2008/1522/pdf/08131.SuJian.ExtAbstrac.1522.pdf.

Usonienė, Aurelija, and Jolė Linkevičienė. 2009. Lietuvių mokslo kalbos tekstynas ir specialioji leksika. *Lituanistica*, 55 (3–4): 133–143. http://mokslozurnalai.lmaleidykla.lt/publ/0235-716X/2009/3-4/133-143.pdf.