Title	Overview of literary corpora in the CLARIN infrastructure
Version	0.2
Author(s)	Darja Fišer, Jakob Lenardič
Date	13-09-2019
Status	Draft
Distribution	BoD, NCF, UI
ID	CE-2019-1517

Table of contents

1. Introduction

In the following report, we present an overview of literary corpora that are part of the CLARIN infrastructure (i.e., they are either listed in the VLO or in the repositories of the national consortia). In this survey, we use the restricted definition of *literature* which comprise poetry and fictional prose (such as novels, short stories and plays) but not non-fiction (such as essays, biographies and religious texts).

The report was conducted in two steps:

(i) by manual searching the VLO and the national consortia with the following keywords: “literature corpus”, “literary corpus”, “poem corpus”, “poetry corpus”, “novel corpus”, “fiction corpus”, “drama corpus”; and
(ii) with input provided by CLARIN UI and NC coordinators.

The full results are available in this Google Docs Spreadsheet. In total, 42 literary corpora were identified. Information on most of the corpora was provided by UI and NC coordinators, whom we would like to thank for their invaluable input. In Section 2, we provide a comprehensive list of the literary corpora that are part of the CLARIN infrastructure, with the corpora divided into monolingual and multilingual resources. In Section 3, we describe their identification (i.e., listed in the VLO or not), their availability (download or through a concordancer), and their metadata (language, size, annotation, license).

2. Literary corpora in the CLARIN infrastructure

2.1. Monolingual literary corpora

The following table lists 34 monolingual literary corpora that are part of the CLARIN infrastructure.

Corpus	Language	Description
One-million Corpus of Croatian Literary Language	Croatian	This corpus is unavailable because the Project URL is broken.
Johannes V. Jensen Corpus **Size:** 1,760,093 words; 8,489 pages **Licence:** CC BY-SA 4.0	Danish	This corpus presents the collected works of the Danish author Johannes Jensen. The corpus is available for download from CLARIN-DK and for online browsing through a dedicated concordancer.
Complete Corpus of Anglo-Saxon Poetry **Annotation:** none	English (Old)	The corpus is available for online browsing through an external interface.
York-Helsinki Parsed Corpus of Old English Poetry **Size:** 71,490 words **Annotation:** MSD-tagged, syntactically parsed **Licence:** Restricted	English (Old)	This corpus contains a selection of poetic texts (71,490 words) from the Old English Section of the Helsinki Corpus of English Texts. The corpus is available for download from the Oxford Text Archive.
Collection of older original Estonian-language works of fiction **Size:** 173 texts **Licence:** CLARIN ACA	Estonian	This corpus collects older Estonian literary texts published on "Kreutzwald's Century: the Estonian Cultural History Web". The electronically republished books, included in the collection, are based on the first editions of works by more important Estonian authors, published in 1854-1944. The corpus is available for online browsing through an external interface.
Corpus of Estonian fiction **Size:** 5,768,504 words **Licence:** CLARIN ACA - NC	Estonian	The corpus contains texts from 1990 (**onwards?**). The corpus is available for download from CELR.
Estonian Runic Songs' Database **Size:** 92,134 texts **Licence:** CLARIN ACA	Estonian	These are the oldest text recordings of Estonian runic songs (the text recordings were created in the 19th century and in the first decades of the 20th century). In addition to the runic songs, the database also has songs of

Commented [MOU1]: pls check

		transitional form and end-rhymed songs (about 6000). The corpus is available for online browsing through an external interface.
Classics of English and American Literature in Finnish (CEAL) **Size:** 3 novels, 484,010 tokens **Annotation:** MSD-tagged, syntactically parsed **Licence:** CLARIN RES + NC	Finnish	The corpus contains Finnish translations of the following three texts: Jane Austen: *Ylpeys ja ennakkoluulo* (*Pride and Prejudice*), translated by Kersti Juva, Teos 2013; Henry James: *Washingtonin aukio* (*Washington Square*), translated by Kersti Juva, Otava 2003; Charles Dickens: *Kolea talo* (*Bleak House*), translated by Kersti Juva, Tammi, 2006. The corpus is available for download through FIN-CLARIN and for online browsing through Korp.
Classics of Finnish Literature, Kielipankki Version **Size:** 1,500,000 words **Licence:** EUPL v.1.1 SA	Finnish	This corpus contains prose fiction, plays, poetry and aphorisms (some written originally in Swedish) of established Finnish authors published from 1880s to 1949. The corpus is available for online browsing through Korp.
Corpus of Early Literary Finnish	Finnish	This corpus is not straightforwardly available, because the Project URL seems to be incorrect.
Corpus of Finnish Literary Classics **Size:** 1,456,658 words	Finnish	This corpus contains works by established Finnish fiction writers from the 1880s to the 1930s. There are different types of prose and plays, as well as lyrics and aphorisms. This corpus is available for online browsing through an external interface.
Corpus of Old Literary Finnish **Size:** 3,428,618 words	Finnish	This corpus contains works of various fields published during the Swedish rule (from the 16th century to about 1810), extensive manuscripts from that period (most of which were later printed), as well as individual almanac and decree texts, sermons and poetry.

		This corpus is available for online browsing through an external interface.
Finnish Corpus (Literature) (UHLCS) **Size:** 68,425 words **Annotation:** “tagged” **Licence:** CLARIN RES	Finnish	This corpus contains samples of Finnish literature published by the WSOY publishing company in the 1990. The corpus is available online through FIN-CLARIN.
The Finnish Gutenberg Corpus **Size:** 34,487,420 words **Licence:** CC-BY	Finnish	The corpus contains Finnish books made available by the Gutenberg project. The texts have not been linguistically annotated. The corpus is available for online browsing through Korp.
The Morpho-Syntactic Database of Mikael Agricola's Works **Size:** 83,678 sentences; 428,314 tokens; 38,308 words **Annotation:** MSD-tagged, syntactically parsed **Licence:** CC-BY-ND	Finnish	This corpus contains the Finnish parts of Mikael Agricola’s works (*Abckiria, Rukouskiria, Se Wsi testamenti, Käsikiria, Messu, Piina, Psaltari, Veisut, Profeetat*). The corpus is available for online browsing through Korp.
République-Bastille (1948-1949) **Size:** 37,965 words **Licence:** CC-BY	French	This corpus contains *République-Bastille*, a novel by Melpo Axioti. This French text is of particular linguistic interest since it is a text written in a language other than the mother tongue and is suited for research on bilingualism and self-translation. It would be worth measuring the naturalness of the language with computational tools, for example. The corpus is available for download from clarin:el.
Greek Medieval Texts **Size:** 3,419,553 words **Licence:** CC-BY-NC	Greek (Ancient), Greek (Modern)	The corpus contains medieval texts contains written material covering the period from the 4th till the 16th century A.D. The texts can be classified into the following categories: religious, poetical-literary, political-historical, hymns, epigrams. The corpus is available for download from clarin:el.
Latvian literature classics	Latvian	This corpus presents classics from the end of the 19th century to the beginning of the 20th century.

		The corpus is unavailable, as the link to the external landing page is broken.
North Saami Corpus (Literature) (UHLCS) **Size:** 17,830 words **Licence:** CLARIN RES +NC +NORED +PLAN	North Sami	This corpus contains Kerttu Vuolab's novel *Cheppari cháráhus*. The corpus is available for online browsing through Korp.
NorGramBank – Fiction in Norwegian Bokmål **Size:** 26,903,637 words; 2,469,916 sentences **Annotation:** syntactically parsed **Licence:** CLARIN ACA	Norwegian (Bokmal)	The corpus, which is based on OCR data from the National Library of Norway, is available for online browsing through INESS.
NorGramBank children's fiction in Norwegian Bokmål **Size:** 4,111,213 words; 389,564 sentences **Annotation:** syntactically parsed **Licence:** CLARIN ACA	Norwegian (Bokmal)	The corpus, which is based on OCR data from the National Library of Norway, is available for online browsing through INESS.
NorGrambank children's fiction in Norwegian Nynorsk **Size:** 1,043,260 words; 106,434 sentences	Norwegian (Nynorsk)	The corpus, which is based on OCR data from the National Library of Norway, is available for online browsing through INESS.
NorGramBank fiction in Norwegian Nynorsk **Size:** 2,884,376 words; 260,285 sentences **Annotation:** syntactically parsed **Licence:** CLARIN ACA	Norwegian (Nynorsk)	The corpus, which is based on OCR data from the National Library of Norway, is available for online browsing through INESS.
1000 Novels Corpus **Size:** 1000 texts **Licence:** CC-BY 4.0	Polish	The corpus is available for download from CLARIN-PL.
1000PLUS Novels Corpus (1.0) **Size:** 1000 texts; 17,352,826 words **Licence:** CC-BY-SA 3.0	Polish	The corpus is available for download from CLARIN-PL.
Late 19th- and Early 20th-Century Polish Novels **Licence:** CC-BY 3.0	Polish	The corpus is available for download from CLARIN-PL.
POE: Microcorpus of 20th century Polish poetry **Licence:** plWordNet	Polish	The corpus is available for download from CLARIN-PL.
LT Corpus **Size:** 1,781,083 words	Portuguese	This corpus contains 70 copyright-free classics (61 Portugal and 9 Brazil) published before 1940.

Licence: CLARIN RES		The corpus is available for download from PORTULAN.
Banco de Datos de Once Novelas Españolas 1951—1971 (SOL) (2014-10-08) **Size:** 1,267,391 tokens; 69,270 sentences **Annotation:** sentence scrambled **Licence:** CC-BY 4.0	Spanish	The corpus is available for download from SWE-CLARIN and for online browsing through Korp.
Electronic corpus of 15th-century Castilian cancionero manuscripts	Spanish	This is a lyric corpus of 15th century cancioneros. The corpus is available for online browsing through an external interface.
Electronic text corpus of Sumerian literature (ETCSL) **Size:** 400 literary compositions	Sumerian	This corpus presents a selection of nearly 400 literary compositions recorded on sources which come from ancient Mesopotamia and date to the late third and early second millennia BCE. The corpus is available for online browsing through an external interface.
August Strindberg's novels (2017-10-16) **Size:** 4,309,037 tokens; 321,759 sentences **Annotation:** sentence scrambling **Licence:** CC-BY 4.0	Swedish	This corpus presents the collected works of August Strindberg. The corpus is available for download from SWE-CLARIN and for online browsing through Korp.
Bonnier novels I (1976/77) (2017-10-04) **Size:** 6,578,675 tokens; 462,625 sentences **Annotation:** sentence scrambling **Licence:** CC-BY 4.0	Swedish	This corpus presents 69 Bonnier novels from 1976-77. The corpus is available for download from SWE-CLARIN and for online browsing through Korp.
Bonnier novels II (1980/81) (2017-03-17) **Size:** 4,304,271 tokens; 298,361 sentences **Annotation:** sentence scrambling **Licence:** CC-BY 4.0	Swedish	This corpus presents 60 Bonnier novels from 1980-81. The corpus is available for download from SWE-CLARIN and for online browsing through Korp.

Commented [MOU2]: pls clarify, a single author/bonnier family/bonnier publisher?

Commented [MOU3]: same as above

2.2. Multilingual literary corpora

The following table lists 8 multilingual literary corpora that are part of the CLARIN infrastructure.

Corpus	Language	Description
MULTEXT-East "1984" annotated corpus 4.0	Bulgarian, Czech, English, Estonian, Hungarian, Macedonian,	This is Parallel corpus of George Orwell's *1984* and its translations.

Size: 12 texts; 79,718 sentences; 1,064,424 words **Annotation:** sentence-alignment, MSD tagging **Licence:** CC BY-NC SA 4.0	Persian, Polish, Romanian, Serbian, Slovak, Slovenian	The corpus is available for download from CLARIN.SI
Anthology of Middle English texts / Santiago Gonzalez y Fernandez-Corugedo **Size:** 4,000 words **Annotation:** none **Licence.** Oxford Text Archive Licence	English (Middle), Hebrew	The corpus contains literary texts from 1100 to 1400. The corpus is available for download from the Oxford Text Archive.
Finnish Folk Poetry **Size:** 7.1 million words **Licence:** CC-BY-NC	Finnish, Karelian, Ludian, Latin, Swedish, Olonets, Izhorian, Votic	The corpus contains poems from 1564 to 1939. The corpus is available for online browsing through Korp.
ParFin 2016, Finnish-Russian Parallel Corpus of Literary Texts **Size:** 2,044,172 tokens **Annotation:** MSD-tagged, syntactically parsed **Licence:** CLARIN RES +NC +INF +ND	Finnish, Russian	The corpus contains Finnish literary texts from 1990-2010 and their translations into Russian aligned at sentence level. The corpus is available for online browsing through Korp.
ParRus 2016, Russian-Finnish Parallel Corpus of Literary Texts **Size:** 5,900,000 tokens **Annotation:** MSD-tagged, syntactically parsed **Licence:** CLARIN RES +NC +INF +ND	Finnish, Russian	The corpus contains Russian literary texts (classical literature & 20th century) and their translations into Finnish aligned at paragraph level. The corpus is available for online browsing through Korp.
Aleksis Kivi Corpus (SKS) **Size:** 413,735 words **Annotation:** MSD-tagged, syntactically parsed **Licence:** CC-BY-NC	Finnish, Swedish	The corpus contains all the known letters, manuscripts and published works by Finnish author Aleksis Kivi (1834–1872). Most of the texts were written in Finnish while some of the letters and manuscripts are in Swedish. The time coverage of the texts: 1855-1871. The corpus is available for online browsing through Korp.
Classics Library of the National Library of Finland - Kielipankki version **Licence:** CC-BY	Finnish, Swedish	The corpus contains literary texts from 1549 to 1944. The corpus is available for online browsing through Korp.
aformes **Size:** 376,250 words **Licence:** CC-BY-NC	Greek (Modern), English	This corpus contains fiction texts from a journal of undergraduate creative writing at the Faculty of English Language and Literature. The corpus is available for download from clarin:el.

3. Overview of the literary corpora in the CLARIN infrastructure

3.1. Identification

There are 42 literary corpora in the CLARIN infrastructure. All of the corpora can be found in the VLO except for the following three:

(1) aformes
(2) République-Bastille (1948-1949)
(3) Greek Medieval Texts

3.2. Availability

In this section, we list the availability of the literary corpora. In the parentheses, we list information on the repository from where the corpus can be downloaded and the interface for online browsing, if applicable.

3.2.1. For online browsing and download

The following 7 corpora are available for online browsing and download:

(1) Johannes V. Jensen Corpus (CLARIN-DK; external interface)
(2) Classics of English and American Literature in Finnish (CEAL) (FIN-CLARIN; Korp)
(3) Classics of English and American Literature in Finnish (CEAL) (FIN-CLARIN; Korp)
(4) Classics of Finnish Literature, Kielipankki Version (FIN-CLARIN; Korp)
(5) Banco de Datos de Once Novelas Españolas 1951—1971 (SOL) (2014-10-08) (SWE-CLARIN; Korp)
(6) Bonnier novels I (1976/77) (2017-10-04) (SWE-CLARIN; Korp)
(7) Bonnier novels II (1980/81) (2017-03-17) (SWE-CLARIN; Korp)

3.2.2. For download

The following 12 corpora are available for download:

(1) York-Helsinki Parsed Corpus of Old English Poetry (CLARIN UK: Oxford Text Archive)
(2) Corpus of Estonian fiction (CELR)
(3) République-Bastille (1948-1949) (clarin:el)
(4) 1000 Novels Corpus (CLARIN-PL)
(5) 1000PLUS Novels Corpus (1.0) (CLARIN-PL)
(6) Late 19th- and Early 20th-Century Polish Novels (CLARIN-PL)
(7) POE: Microcorpus of 20th century Polish poetry (CLARIN-PL)
(8) LT Corpus (PORTULAN)
(9) MULTEXT-East "1984" annotated corpus 4.0 (CLARIN.SI)
(10) Anthology of Middle English texts / Santiago Gonzalez y Fernandez-Corugedo (CLARIN UK: Oxford Text Archive)
(11) Greek Medieval Texts (clarin:el)
(12) aformes (clarin:el)

3.2.3. For online browsing

The following 21 corpora are available for online browsing:

(1) Complete Corpus of Anglo-Saxon Poetry (External interface)
(2) Collection of older original Estonian-language works of fiction (External interface)
(3) Estonian Runic Songs' Database (External interface)
(4) Finnish Corpus (Literature) (UHLCS) (FIN-CLARIN)
(5) The Finnish Gutenberg Corpus (Korp)
(6) The Morpho-Syntactic Database of Mikael Agricola's Works (Korp)
(7) North Saami Corpus (Literature) (UHLCS) (Korp)
(8) NorGramBank – Fiction in Norwegian Bokmål (INESS)
(9) NorGramBank children’s fiction in Norwegian Bokmål (INESS)
(10) NorGrambank children's fiction in Norwegian Nynorsk (INESS)
(11) NorGramBank fiction in Norwegian Nynorsk (INESS)
(12) Electronic corpus of 15th-century Castilian cancionero manuscripts (External interface)
(13) Electronic text corpus of Sumerian literature (ETCSL) (External interface)
(14) August Strindberg's novels (2017-10-16) (Korp)

(15) Finnish Folk Poetry (Korp)
(16) ParFin 2016, Finnish-Russian Parallel Corpus of Literary Texts (Korp)
(17) ParRus 2016, Russian-Finnish Parallel Corpus of Literary Texts (Korp)
(18) Aleksis Kivi Corpus (SKS) (Korp)
(19) Classics Library of the National Library of Finland - Kielipankki version (Korp)
(20) Corpus of Old Literary Finnish (External interface)
(21) Corpus of Finnish Literary Classics (External interface)

3.2.4. Unavailable

(1) Latvian literature classics (broken link to landing page)
(2) Corpus of Early Literary Finnish (incorrect link to landing page)
(3) One-million Corpus of Croatian Literary Language (broken link to landing page)

3.3. Metadata

3.3.1. Language

There are 34 monolingual literary corpora in the CLARIN infrastructure, accounting for the following 15 languages:

(1) Croatian (1 corpus)
(2) Danish (1 corpus)
(3) English (2 corpora)
(4) Estonian (3 corpora)
(5) Finnish (8 corpora)
(6) French (1 corpus)
(7) Greek (1 corpus)
(8) Latvian (1 corpus)
(9) North Sami (1 corpus)
(10) Norwegian (4 corpora)
(11) Polish (4 corpora)
(12) Portuguese (1 corpus)
(13) Spanish (2 corpora)
(14) Sumerian (1 corpus)
(15) Swedish (3 corpora)

There are 8 multilingual corpora, accounting for the following 6 language combinations:

(1) Bulgarian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Romanian, Serbian, Slovak, Slovenian (1 corpus)
(2) English (Middle), Hebrew (1 corpus)
(3) Finnish, Karelian, Ludian, Latin, Swedish, Olonets, Izhorian, Votic (1 corpus)
(4) Finnish, Russian (2 corpora)
(5) Finnish, Swedish (2 corpora)
(6) Greek, English (1 corpus)

3.3.2. Size

Information on size is generally readily included (available for approx. 83% of the corpora); only the following 7 corpora lack this information:

(1) Complete Corpus of Anglo-Saxon Poetry
(2) Latvian literature classics
(3) Late 19th- and Early 20th-Century Polish Novels
(4) POE: Microcorpus of 20th century Polish poetry
(5) Electronic corpus of 15th-century Castilian cancionero manuscripts
(6) Classics Library of the National Library of Finland - Kielipankki version
(7) Corpus of Early Literary Finnish

This is the size overview for corpora that contain information on word/token-number:

- 8 corpora contain <1 million words/tokens
- 16 corpora contain 1–10 million words/tokens
- 2 corpora contain >10 million words/tokens

It is worth noting that not all corpora report size in terms of words/tokens; many simply state the number of texts included:

(1) Collection of older original Estonian-language works of fiction (173 texts)
(2) Estonian Runic Songs' Database (92,134 texts)
(3) 1000 Novels Corpus (1000 texts)
(4) Electronic text corpus of Sumerian literature (ETCSL) (400 literary compositions)

By contrast, quite a few corpora include a more fine-grained presentation on size: e.g., the syntactically annotated NorGramBank – Fiction in Norwegian Bokmål corpus reports on both word-size and sentence-size, while MULTEXT-East "1984" annotated corpus 4.0 reports on the number of texts, sentence size, and word size.

3.3.3. Annotation

Information on annotation is available for less than half of all the corpora (19 out of 42):

(1) Johannes V. Jensen Corpus
(2) Collection of older original Estonian-language works of fiction
(3) Corpus of Estonian fiction
(4) Estonian Runic Songs' Database
(5) Classics of Finnish Literature, Kielipankki Version
(6) The Finnish Gutenberg Corpus
(7) République-Bastille (1948-1949)
(8) Latvian literature classics
(9) North Saami Corpus (Literature) (UHLCS)
(10) 1000 Novels Corpus
(11) 1000PLUS Novels Corpus (1.0)
(12) Late 19th- and Early 20th-Century Polish Novels
(13) POE: Microcorpus of 20th century Polish poetry
(14) LT Corpus
(15) Electronic corpus of 15th-century Castilian cancionero manuscripts
(16) Electronic text corpus of Sumerian literature (ETCSL)
(17) Classics Library of the National Library of Finland - Kielipankki version
(18) Greek Medieval Texts
(19) aformes

In certain cases, this information is omitted because the corpus lacks any kind of linguistic annotation. For instance, both Electronic corpus of 15th-century Castilian cancionero manuscripts and Estonian Runic Songs' Database lack this information, but only the first truly is unannotated (it is a collection of scanned manuscripts and their transcriptions), whereas the second is PoS-tagged which is undocumented. We believe that it is vital to explicitly state in the metadata if a corpus has not been annotated; e.g., Complete Corpus of Anglo-Saxon Poetry is described as a “plain-text” corpus, which helpfully points out its lack of annotation.

Otherwise, the following annotation layers are included:

- MSD-tagging (10 corpora)
- Syntactic parsing (10 corpora)
- Sentence scrambling (4 corpora)

3.3.4. Licence

Information on licence is missing for the following 8 corpora:

(1) Complete Corpus of Anglo-Saxon Poetry

(2) Latvian literature classics
(3) Electronic corpus of 15th-century Castilian cancionero manuscripts
(4) Electronic text corpus of Sumerian literature (ETCSL)
(5) Corpus of Finnish Literary Classics
(6) Corpus of Early Literary Finnish
(7) Corpus of Old Literary Finnish
(8) One-million Corpus of Croatian Literary Language

The most common licence types are the following:

- CC-BY (17 corpora)
- CLARIN RES (7 corpora)
- CLARIN ACA (6 corpora)

4. Conclusion

In this report, we presented an overview of 42 literary corpora in the CLARIN infrastructure. We presented their identification (i.e., whether they have VLO entries) and their availability (for download, online browsing or both), as well as 4 types of metadata – language, size, annotation, and licence.

In terms of identification, 3 (7.1%) out of the 42 identified corpora are not listed in the VLO, all of which belong to the *clarin:el* repository. In terms of availability, 3 (7.1%) out of the 42 corpora are unavailable either for download or online browsing because the link to its external landing page is broken in 2 cases, while 1 corpus seems to have an incorrect Project URL, leading to an incorrect corpus. Otherwise, availability is as follows: 7 (16.6%) corpora are available both for online browsing and download, 12 (28.6%) corpora are available only for download, and 21 (50%) corpora are available only for online browsing. It is noteworthy that out of the 28 corpora that can be browsed online, most (20 or 71.4%) are available through the concordancer Korp (FIN-CLARIN or SWE-CLARIN distribution), while 8 (28.6%) are available through external interfaces.

Most corpora cover languages spoken in Europe, with Finnish being the most represented language (8 or 19% of the 42 corpora). It is noteworthy that there is also a corpus of literary compositions in ancient Sumerian, which is a language isolate and one of the oldest languages for which exist written samples.

Information on size is mostly included, and is only missing for 7 (16.7%) out of the 42 corpora. An issue is unequal metadata granularity – there exist corpora which report different types of size (number of texts, number of tokens, number of sentences), while some corpora only report the number of texts included, even though they seem to be tokenised.

Information on annotation fares slightly worse, as it is included for less than half of the 42 corpora. For literary corpora, including this information is especially crucial because there exist quite a few plain-text resources, so it can be unclear whether the metadata of a corpus are simply insufficiently described or the corpus is not annotated. Especially in the domain of literature, the term *corpus* appears to be often used in its most relaxed sense denoting simply a collection of often unannotated literary texts the comprehensiveness or sampling criteria are frequently undocumented. Therefore, our main proposal is that creators/curators of literary corpora (as well as corpora in other domains) always make explicit the annotation process in the documentation, even in case the corpus is simply a collection of plain texts.

Finally, information on licence is readily included type, and is missing for only 8 (19.1%) out of the 42 corpora. 17 (40.5%) corpora are available under CC-BY, 7 (16.7%) under CLARIN RES, 6 (14.3%) under CLARIN ACA, and the rest under miscellaneous licences.