Title	Overview of historical corpora
Version	1.0
Author(s)	Darja Fišer, Jakob Lenardič
Date	06-06-2018
Status	For distribution
Distribution	BoD, NCF, UI
ID	CE-2018-1237

Table of contents

1. Introduction

In the following report, we present an overview of historical corpora, primarily focusing 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).

The report was conducted in two steps:

(i) manually searching the VLO and the national consortia with keywords like "historic* corpus", "medieval corpus", etc.
(ii) input provided by CLARIN UI and NC coordinators

The full results are available in a Google Docs Spreadsheet.[1] In total, around 90 historical 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 historical corpora that are part of the CLARIN infrastructure, describing their identification (i.e., listed in the VLO or not), their availability (download or through a concordancer), and their metadata (language, size, annotation, license). In section 3, we provide a list of historical corpora are available outside the CLARIN infrastructure. In

[1] https://docs.google.com/spreadsheets/d/1orDix-eyoHE3K2H44HVNs3ttA_SmdDwzCrFKtcJPV8E/edit

section 4, we provide a short list of tools and services that were provided by participants in the Google forms survey.

2. Historical corpora in the CLARIN infrastructure

Table 1 lists 68 historical corpora that are part of the CLARIN infrastructure

Table 1: historical corpora in the CLARIN infrastructure

Corpus	Language	Description
Open Richly Annotated Cuneiform Corpus, Korp Version **Size:** 741,100 tokens **Annotation:** tokenised **Licence:** CC-BY-SA	Akkadian	This corpus contains cuneiform texts from Ancient history. The corpus is available through the concordancer Korp.
Greek Medieval Texts **Size:** 3.4 million words **Licence:** CC-BY	Ancient Greek	This corpus contains texts from the 4th to the 16th century. The corpus is available for download from the clarin:el repository.
Sheffield Corpus of Chinese **Licence:** CC-BY-NC-SA 3.0	Chinese	This corpus contains fictional and non-fictional texts from the Medieval and Modern Chinese periods. The corpus is available for download from the Oxford Text Archive.
"PolDiLemma" Middle Polish Diachrone Lemmatised Corpus **Annotation:** tokenised, lemmatised	Czech, German, Latin, Polish	This corpus contains political, religious and scientific texts from the 16th to the 18th century. The corpus is available for download from the CLARIN-D repository.
Medieval Charter Sections Corpus **Size:** 57 chapters **Annotation:** manually-tagged, named entities **Licence:** CC-BY-NC-SA 4.0	Czech, Latin	This corpus contains Latin charters created in the era of John the Bling, King of Bohemia. The corpus is available for download from LINDAT.
Brieven als buit (Letters as loot) **Size:** 460,000 words **Annotation:** lemmatised, PoS-tagged, grammatically tagged	Dutch	This corpus contains 1,000 letters from the 17th to the 18th century. The corpus is available through a dedicated concordancer.
Corpus Gysseling **Size:** 1.5 million words **Annotation:** PoS-tagged, lemmatised **Licence:** INT Licence for researchers	Dutch	This corpus contains texts from the 13th century. The corpus is available for download from the Instituut voor de Nederlandse Taal and through a dedicated concordancer.

A Corpus of English Dialogues 1560-1760 (CED) **Size:** 1.2 million words **Licence:** Oxford Text Archive licence	English	This corpus contains dialogues from literary and didactic works from 1560 to 1760. The corpus is available for download from the Oxford Text Archive.
Corpus of Early English Correspondence Sampler (CEECS) **Size:** 450,000 words **Annotation:** none **Licence:** Oxford Text Archive licence	English	This corpus contains 1147 letters from 1418 to 1680. The corpus is available for download from the Oxford Text Archive.
Corpus of Late Modern English prose / David Denison **Annotation:** none **Licence**: Oxford Text Archive licence	English	This corpus contains fictional texts from 1837 to 1926. The corpus is available for download from the Oxford Text Archive.
Hansard Corpus **Size:** 1.6 billion tokens **Annotation:** tokenised, PoS-tagged, lemmatised, semantic tags	English	This corpus contains parliamentary debates from 1803 to 2005. The corpus is available through a dedicated concordancer.
Helsinki Corpus of Scottish Correspondence (1540-1750) **Size:** 0.5 million tokens **Annotation:** tokenised **Licence:** CLARIN ACA	English	This corpus contains personal correspondence from 1540 to 1750. The corpus is available through the concordancer Korp.
Older Scottish texts : the Edinburgh DOST corpus / A.J. Aitken, Paul Bratley and Neil Hamilton-Smith **Size:** 877,000 tokens **Annotation:** tokenised **Licence:** CC-BY-NC-SA 3.0	English	This corpus contains texts from 1450 to 1600. The corpus is available for download from the Oxford Text Archive.
Pamphlets of the American Revolution : [selections] / edited by Bernard Bailyn **Annotation:** none **Licence**: CC-BY-NC-SA 3.0	English	This corpus contains pamphlets of the American Revolution from 1750 to 1776. The corpus is available for download from the Oxford Text Archive.

Parsed Corpus of Early English Correspondence (PCEEC) **Size:** 2.2 million words **Annotation:** tokenised, PoS-tagged, syntactically parsed **Licence:** Oxford Text Archive licence	English	This corpus contains correspondence from around 1410 to 1681. This corpus is available for download from the Oxford Text Archive.
The English language of the north-west in the late Modern English period: a Corpus of late 18c Prose **Size:** 300,000 words **Annotation:** COCOA-style **Licence**: Oxford Text Archive licence	English	This corpus contains texts from 1761 to 190. The corpus is available for download from the Oxford Text Archive.
The Lampeter Corpus of Early Modern English Tracts **Annotation:** none **Licence**: CC-BY-NC-SA 3.0	English	This corpus contains tracts from 1640 to 1740. The corpus is available for download from the Oxford Text Archive.
The Lancaster Newsbooks Corpus **Licence:** CC-BY-NC-SA 3.0	English	This corpus contains two collections of English newsbooks from 1654 to 1655. The corpus is available for download from the Oxford Text Archive.
Corpus of Historical American English - Kielipankki Korp version 2017H1 **Size:** 385 million tokens **Annotation:** tokenised **Licence:** CLARN ACA	English (American)	This corpus contains texts from 1810 to 2009. The corpus is available through the concordancer *Korp*.
The Old Bailey Corpus **Size:** 134 million words **Annotation:** detailed sociobiographical, pragmatic and textual annotation **Licence:** CC-BY-NC-SA 4.0	English (Late Modern)	This corpus contains proceedings of the Old Bailey (i.e., legal documents) from 1674 to 1913. The corpus is available for download from the CLARIN-D repository and through the CQPConcordancer.
Anthology of Middle English texts / Santiago Gonzalez y Fernandez-Corugedo **Size:** 4000 words **Annotation:** none **Licence:** Oxford Text Archive licence	English (Middle), Hebrew	This corpus contains literary texts from 1100 to 1400. The corpus is available for download from the Oxford Text Archive.

Helsinki corpus of English texts **Size:** 240,000 words **Annotation:** none **Licence**: Oxford Text Archive licence	English (Old and Middle)	This corpus contains Biblical and fictional texts from 730 to 1710. The corpus is available for download from the Oxford Text Archive.
The York-Helsinki parsed corpus of Old English poetry (YCOEP) **Size:** 71,500 words **Annotation:** syntactically-parsed **Licence**: Oxford Text Archive licence	English (Old)	This corpus contains poems from 730 to 1710. The corpus is available for download from the Oxford Text Archive.
Dictionary of Old English Corpus in Electronic Form (DOEC) **Licence:** Oxford Text Archive licence	English (Old), Latin	This corpus contains 3037 texts from 600 to 1150. The corpus is available for download from the Oxford Text Archive.
The York-Toronto-Helsinki Parsed Corpus of Old English prose (YCOE) **Size:** 1.5 million words **Annotation:** syntactically-parsed **Licence**: Oxford Text Archive licence	English (Old), Latin	This corpus contains fictional texts from 600 to 1150. The corpus is available for download from the Oxford Text Archive.
Hamburg Corpus of Old Swedish with Syntactic Annotations (HaCOSSA) **Size:** 128,000 words **Annotation:** MSD-tagged, syntactically parsed **Licence:** CLARIN RES	English, German, Latin, Old Norse, Swedish	This corpus contains texts written in the Late Old Swedish period (from 1375 to 1550). The corpus is available for download from the repository of the University of Hamburg.
The Electronic Text Corpus of Sumerian Literature. Revised edition. **Annotation:** Each word form in the composite transliterations has been assigned to a lexeme which is specified by a citation form, word class information and basic English translation. **Licence:** CC-BY-NC-SA 3.0	English, Sumerian	This corpus contains transliterations and English translations of 394 Sumerian compositions from approximately 2100 to 1700 BCE. The corpus is available for download from the Oxford Text Archive.

Corpus of Old Written Estonian **Size:** 2 million tokens **Annotation:** tokenised, 16.-18. century texts have been tagged with contemporary Estonian, morphological and language information. 19. century texts are unannotated. **Licence:** CC-BY	Estonian	This corpus covers secular and religious texts from the 16th to the 18th century. The corpus is available through a dedicated concordancer.
Classics of Finnish Literature, Kielipankki Version **Size:** 1.5 million words **Licence:** EUPL v.1.1 SA	Finnish	This corpus contains literary texts from 1880 to 1949. The corpus is available through the concordancer Korp.
Corpus of Old Literary Finnish **Size:** 4.1 million words **Licence:** EUPL v.1.1 SA	Finnish	This corpus contains literary texts from 1543 to 1810. The corpus is available through the concordancer Korp.
The Finnish Gutenberg Corpus **Size:** 34.5 million words **Annotation:** no linguistic annotation **Licence:** CC-BY	Finnish	This corpus contains books published up to 1925 that are made available through the Gutenberg project. The corpus is available through the concordancer Korp.
The Finnish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version **Size:** 5.2 billion tokens **Annotation:** tokenised **Licence:** CC-BY-SA	Finnish	This corpus contains newspaper articles from 1840 to 2011. The corpus is available through the concordancer Korp.
The Morpho-Syntactic Database of Mikael Agricola's Works **Size:** 428,300 tokens **Annotation:** tokenised, PoS-tagged, morphological components and syntactic function **Licence:** CC-BY-ND	Finnish	This corpus contains texts from 1544 to 1551 written by the clergyman Mikael Agricola. The corpus is available through the concordancer Korp.
Virtual Old Literary Finnish (VVKS) - Kielipankki Korp version **Size:** 48 texts **Licence:** CC-BY-NC-ND	Finnish	This corpus will contain literary texts from 1543 to 1791.

Finnish Folk Poetry **Size:** 7.1 million words **Licence:** CC-BY-NC	Finnish, Karelian, Ludian, Latin, Swedish, Olonets, Izhorian, Votic	This corpus contains poems from 1564 to 1939. The corpus is available through the concordancer Korp.
Corpus of Early Modern Finnish, Kielipankki Version **Size:** 8.6 million words **Licence:** EUPL v.1.1 SA	Finnish, Russian, German, Latin	This corpus contains texts from 1809 to 1899. The corpus is available through the concordancer Korp.
Aleksis Kivi Corpus (SKS) **Size:** 413,700 words **Licence:** CC-BY-NC	Finnish, Swedish	This corpus contains the works by Finnish author Aleksis Kivi from 1855 to 1871. The corpus is available through the concordancer Korp.
Classics Library of the National Library of Finland - Kielipankki version **Licence:** CC-BY	Finnish, Swedish	This corpus will contain literary texts from 1549 to 1944.
The Letters of Paul Sinebrychoff, Kielipankki Version **Size:** 8.6 million words **Licence:** CC-BY	Finnish, Swedish	This corpus contains letters from 1895 to 1909. The corpus is available through the concordancer Korp.
The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version **Size:** 8.7 billion words **Licence:** CC-BY	Finnish, Swedish	This corpus contains newspaper articles from 1770 to 2011. The corpus is available through the concordancer Korp.
The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874) **Licence:** CC-BY	Finnish, Swedish	This corpus contains newspaper articles from 1771 to 1874. The corpus is available for download from the Language Bank of Finland.
The Newspaper and Periodical OCR Corpus of the National Library of Finland (1875-1920) **Size:** 8.7 billion tokens **Annotation:** tokenised **Licence:** CLARIN ACA	Finnish, Swedish	This corpus contains newspaper articles from 1875 to 1920. The corpus is available for download from the Language Bank of Finland.
Partonopeus de Blois: transcriptions of all manuscripts and fragments **Licence:** CC-BY-NC-SA 3.0	French (Old)	This corpus contains transcriptions of the manuscripts and fragments of the romance Partonopeus de Blois. The corpus is available for download from the Oxford Text Archive.

Syntactic Reference Corpus of Medieval French **Size:** 245,000 tokens **Annotation:** tokenised, syntactically-parsed **Licence:** CLARIN ACA	French (Old)	This corpus contains texts from the 9th to the 13th century. The corpus is available for download from a dedicated webpage.
Austrian Baroque Corpus **Size:** 200,000 tokens **Annotation:** tokenised, PoS-tagged, lemmatised, named entities	German	This corpus contains sermons from 1650 to 1750. The corpus is available through a dedicated concordancer.
DDR-Presseportal (GDR press portal) **Size:** 1.1 billion tokens **Annotation:** tokenised, lemmatised, PoS-tagged, normalised orthography **Licence:** CLARIN ACA	German	This corpus contains newspaper texts from 1945 to 1994. The corpus is available through a concordancer provided by CLARIN-D.
Deutsches Textarchiv (DTA) **Licence:** CLARIN PUB	German	This corpus contains texts from the 17th to the 20th century.
Die Grenzboten (journal) **Size:** 89 million tokens **Annotation:** tokenised, lemmatised, PoS-tagged, normalised orthography	German	This corpus contains texts from 1842 to 1921. The corpus is available for download from the Deutsches Text Archiv and through a concordancer.
Dinglers Polytechnisches Journal (Polytechnical Journal of Dingler) **Size:** 77.5 million tokens **Annotation:** tokenised, PoS-tagged, lemmatised, normalized orthography **Licence:** CC-BY-NC-SA 3.0	German	This corpus contains academic texts from 1820 to 1931. The corpus is available for download from the Deutsches Text Archiv and through a concordancer.
GeMi Corpus **Size:** 120,000 tokens **Annotation:** TEI Lite markup, no linguistic annotation **Licence:** CC-BY-NC-SA 3.0	German	This corpus contains medical writing from 1500 to 1700. The corpus is available for download from the Oxford Text Archive.
GerManC. A Historical Corpus of German Newspapers 1650-1800 **Size:** 700,000 words **Licence:** CC-BY-NC-SA 3.0	German	This corpus contains personal letters, sermons and fictional, scholarly (i.e., humanities), scientific and legal texts from 1650 to 1800. The corpus is available for download from the Oxford Text Archive.

Mannheimer Korpus Historischer Zeitungen und Zeitschriften **Size:** 3532 pages	German	This corpus contains texts from the 18th and 19th centuries. The corpus is available for download directly through the VLO
Referenzkorpus Mittelhochdeutsch (Middle High German Reference Corpus **Size:** 2.5 million tokens **Annotation:** tokenised, PoS-tagged, lemmatised, normalised, morphosyntactic description **Licence:** CC-BY-SA 4.0	German	This corpus contains texts from 1050 to 1350. The corpus is available for download from the Deutsches Text Archiv and through a concordancer.
B4 Historisches Predigtenkorpus zum Nachfeld **Size:** 92,500 tokens **Annotation:** tokenised, syntactic and discursive annotation **Licence:** CLARIN ACA	German (Middle High)	This corpus contains sermons from an Upper German (Balvarian-Alemannic) dialect area. The corpus is available for download from the repository of the University of Hamburg.
B4 Ludolf **Size:** 6,690 tokens **Annotation:** tokenised, tagged for clause type and grammatical function **Licence:** CLARIN ACA	German (Middle Low)	This corpus contains texts from a journey diary from 1350. The corpus is available for download from the repository of the University of Hamburg.
Reference Corpus Middle Low German/Low Rhenish (1200-1650) **Size:** 200,700 tokens **Annotation:** tokenised, MSD-tagged **Licence:** CC-BY	German (Middle Low)	This corpus contains texts from the 13th century to the middle of the 17th century. The corpus is available for download from the repository of the University of Hamburg.
OROSSIMO Corpus – History **Size:** 553,000 tokens **Annotation:** structural annotation (paragraph) **Licence:** CC-BY	Greek	This corpus contains historic academic texts. The corpus is available for download from the clarin:el repository.
Hungarian Historical Corpus **Size:** 30 million words	Hungarian	This corpus contains historical texts from the 18th century to the 2000s. The corpus is available through a dedicated concordancer.

B4 Tatian Corpus of Deviating Examples 2.1 **Size:** 11,300 tokens **Annotation:** tokenised, MSD-tagged **Licence:** CC-BY	Latin, German (Old High)	This corpus contains the OHG Tatian, which is one of the largest prose texts from the Old High German period. The corpus is available for download and through a concordancer from the repository of the University of Hamburg.
Menota **Size:** 1.6 million tokens **Annotation:** tokenised, MSD-tagged, lemmatised **Licence:** CC-BY	Old Norse	This corpus contains Medieval Nordic texts. The corpus is available through the concordancer *Corpuscle*.
Polish language of the 1960s **Size:** 500,000 words **Annotation:** MSD-**tagged** **Licence:** CC-BY-NC-SA 3.0	Polish	This corpus contains essays, news articles, and scientific and literary texts from 1963 to 1967. The corpus is available for download from the Oxford Text Archive.
Corpus of biblical text in Scots / John Kirk **Annotation:** none **Licence**: Oxford Text Archive licence	Scots	This corpus contains Biblical texts. The corpus is available for download from the Oxford Text Archive.
The Helsinki corpus of Older Scots : [1450-1700] **Annotation:** none **Licence**: CC-BY-NC-SA 3.0	Scots	This corpus contains texts of different domains and genres (e.g., burgh records, diaries, pamphlets, scientific treatises, sermons) from 1450 to 1700. The corpus is available for download from the Oxford Text Archive.
Digital library and corpus of historical Slovene IMP 1.1 **Size:** 17.7 million tokens **Annotation:** tokenised, lemmatised, PoS-tagged **Licence:** CC-BY-SA 4.0	Slovenian	This corpus contains 658 unique texts from 1584 to 1919. The corpus is available for download from the CLARIN.SI repository and through the concordancer KonText.
Reference corpus of historical Slovene goo300k 1.2 **Size:** 300,000 tokens **Annotation:** manually tokenised, lemmatised, PoS-tagged, modern synonyms for archaic words **Licence:** CC-BY 4.0	Slovenian	This corpus contains 89 unique texts from 1584 to 1899. The corpus is available for download from the CLARIN.SI repository and through the concordancer KonText.

The Swedish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version **Size:** 3.5 billion tokens **Annotation:** tokenised **Licence:** CC-BY-SA.	Swedish	This corpus contains newspaper articles from 1770 to 1950. The corpus is available through the concordancer Korp.
Språkbanken's historical corpora **Size:** 1.34 billion tokens **Annotation:** tokenised, PoS-tagged, lemmatised, syntactically parsed, word sense (for materials more recent than 1800)	Swedish, German, French and others	This collection of corpora contains – among others – diachronic legal texts, Bible translations, medieval letters, digitized newspapers from the Swedish National Library and 19th century fiction from the Swedish Literature Bank. The corpora are available through the concordancer Korp.
Historical Corpus of the Welsh Language 1500-1850 **Size:** 420,000 words	Welsh	This corpus contains 30 texts from 1500 to 1850. The corpus is available for download from a dedicated website and through a dedicated concordancer.

2.1. Summary

2.1.1. Identification

56 out of the 68 CLARIN corpora can be found in the VLO. The following 12 corpora are not yet included in the VLO:

1. Menota
2. Greek Medieval Texts
3. Austrian Baroque Corpus
4. OROSSIMO Corpus - History
5. DDR-Presseportal (GDR press portal)
6. The Morpho-Syntactic Database of Mikael Agricola's Works
7. Classics Library of the National Library of Finland - Kielipankki version
8. Virtual Old Literary Finnish (VVKS) - Kielipankki Korp version
9. The Newspaper and Periodical OCR Corpus of the National Library of Finland (1875-1920)
10. The Finnish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
11. The Swedish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
12. Språkbanken's historical corpora

2.1.2. Availability

The following 9 corpora are available for download and through a concordancer:

1. Reference corpus of historical Slovene goo300k 1.2

2. Digital library and corpus of historical Slovene IMP 1.1
3. Historical Corpus of the Welsh Language 1500-1850
4. The Old Bailey Corpus
5. B4 Tatian Corpus of Deviating Examples 2.1
6. Dinglers Polytechnisches Journal (Polytechnical Journal of Dingler)
7. Referenzkorpus Mittelhochdeutsch (Middle High German Reference Corpus
8. Die Grenzboten (journal)
9. Corpus Gysseling

The following 22 corpora are available through a concordancer:

1. Hungarian Historical Corpus
2. Corpus of Historical American English - Kielipankki Korp version 2017H1
3. Helsinki Corpus of Scottish Correspondence (1540-1750)
4. Menota
5. Austrian Baroque Corpus
6. Hansard Corpus
7. DDR-Presseportal (GDR press portal)
8. Brieven als buit (Letters as loot)
9. The Morpho-Syntactic Database of Mikael Agricola's Works
10. The Finnish Gutenberg Corpus
11. Aleksis Kivi Corpus (SKS)
12. Finnish Folk Poetry
13. Classics of Finnish Literature, Kielipankki Version
14. Corpus of Old Literary Finnish
15. Corpus of Early Modern Finnish, Kielipankki Version
16. The Letters of Paul Sinebrychoff, Kielipankki Version
17. The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
18. Open Richly Annotated Cuneiform Corpus, Korp Version
19. The Finnish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
20. The Swedish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
21. Språkbanken's historical corpora
22. Corpus of Old Written Estonian

The following 35 corpora are available for download:

1. Medieval Charter Sections Corpus
2. Sheffield Corpus of Chinese
3. GerManC. A Historical Corpus of German Newspapers 1650-1800
4. "PolDiLemma" Middle Polish Diachrone Lemmatised Corpus
5. "PolDiLemma" Middle Polish Diachrone Lemmatised Corpus
6. Parsed Corpus of Early English Correspondence (PCEEC)
7. Syntactic Reference Corpus of Medieval French
8. Hamburg Corpus of Old Swedish with Syntactic Annotations (HaCOSSA)
9. Reference Corpus Middle Low German/Low Rhenish (1200-1650)

10. B4 Ludolf
11. B4 Historisches Predigtenkorpus zum Nachfeld
12. Mannheimer Korpus Historischer Zeitungen und Zeitschriften
13. Greek Medieval Texts
14. OROSSIMO Corpus – History
15. Older Scottish texts : the Edinburgh DOST corpus / A.J. Aitken, Paul Bratley and Neil Hamilton-Smith
16. Anthology of Middle English texts / Santiago Gonzalez y Fernandez-Corugedo
17. Helsinki corpus of English texts
18. Corpus of biblical text in Scots / John Kirk、
19. Pamphlets of the American Revolution : [selections] / edited by Bernard Bailyn
20. Corpus of Late Modern English prose / David Denison
21. The Helsinki corpus of Older Scots : [1450-1700]
22. The Lampeter Corpus of Early Modern English Tracts
23. The York-Helsinki parsed corpus of Old English poetry (YCOEP)
24. Corpus of Early English Correspondence Sampler (CEECS)
25. The York-Toronto-Helsinki Parsed Corpus of Old English prose (YCOE)
26. The English language of the north-west in the late Modern English period: a Corpus of late 18c Prose
27. Polish language of the 1960s
28. Dictionary of Old English Corpus in Electronic Form (DOEC)
29. Partonopeus de Blois: transcriptions of all manuscripts and fragments
30. A Corpus of English Dialogues 1560-1760 (CED)
31. The Electronic Text Corpus of Sumerian Literature. Revised edition.
32. The Lancaster Newsbooks Corpus
33. GeMi Corpus
34. The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874)
35. The Newspaper and Periodical OCR Corpus of the National Library of Finland (1875-1920)

2.1.3. Metadata

2.1.3.1. Language

There are 51 monolingual historical corpora in the CLARIN infrastructure for the following 16 languages:

1. English (15 corpora)
2. German (12 corpora)
3. Finnish (6 corpora)
4. French (2 corpora)
5. Dutch (2 corpora)
6. Greek (2 corpora)
7. Slovenian (2 corpora)
8. Scots (2 corpora)

9. Akkadian (1 corpus)
10. Chinese (1 corpus)
11. Estonian (1 corpus)
12. Hungarian (1 corpus)
13. Old Norse (1 corpus)
14. Polish (1 corpus)
15. Swedish (1 corpus)
16. Welsh (1 corpus)

The other 17 corpora are multilingual, for the following language combinations:

1. Finnish, Swedish (6 corpora)
2. English, Latin (2 corpora)
3. Czech, German, Latin, Polish (1 corpus)
4. Czech, Latin (1 corpus)
5. English, Hebrew (1 corpus)
6. English, German, Latin, Old Norse, Swedish (1 corpus)
7. English, Sumerian (1 corpus)
8. Finnish, Karelian, Ludian, Latin, Swedish, Olonets, Izhorian, Votic (1 corpus)
9. Finnish, Russian, German, Latin (1 corpus)
10. Latin, German (1 corpus)
11. Swedish, German, French, etc. (1 corpus)

2.1.3.2. Size and temporal span

Information on word/token size is available for 53 corpora. The following 15 corpora lack such information:

1. Sheffield Corpus of Chinese
2. "PolDiLemma" Middle Polish Diachrone Lemmatised Corpus
3. Deutsches Textarchiv (DTA)
4. Corpus of biblical text in Scots / John Kirk
5. Pamphlets of the American Revolution : [selections] / edited by Bernard Bailyn
6. Corpus of Late Modern English prose / David Denison
7. The Helsinki corpus of Older Scots : [1450-1700]
8. The Lampeter Corpus of Early Modern English Tracts
9. Dictionary of Old English Corpus in Electronic Form (DOEC)
10. Partonopeus de Blois: transcriptions of all manuscripts and fragments
11. The Electronic Text Corpus of Sumerian Literature. Revised edition.
12. The Lancaster Newsbooks Corpus
13. Classics Library of the National Library of Finland - Kielipankki version
14. Virtual Old Literary Finnish (VVKS) - Kielipankki Korp version
15. The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874)

The largest corpus is The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version, containing 8.7 billion tokens.

In terms of termporal span, there is no common denominator, but generally the oldest corpora reach back to the 5th century (e.g., there are 3 Old English corpora). The exception is Open Richly Annotated Cuneiform Corpus, Korp Version, which contains Sumerian texts and is thus the oldest corpus by far.

2.1.3.3. Annotation

Information on linguistic annotation is available for 47 corpora. The following 21 corpora lack such information:

1. Hungarian Historical Corpus
2. Sheffield Corpus of Chinese
3. Historical Corpus of the Welsh Language 1500-1850
4. GerManC. A Historical Corpus of German Newspapers 1650-1800
5. Deutsches Textarchiv (DTA)
6. Mannheimer Korpus Historischer Zeitungen und Zeitschriften
7. Greek Medieval Texts
8. Dictionary of Old English Corpus in Electronic Form (DOEC)
9. Partonopeus de Blois: transcriptions of all manuscripts and fragments
10. A Corpus of English Dialogues 1560-1760 (CED)
11. The Lancaster Newsbooks Corpus
12. Aleksis Kivi Corpus (SKS)
13. Finnish Folk Poetry
14. Classics of Finnish Literature, Kielipankki Version
15. Corpus of Old Literary Finnish
16. Corpus of Early Modern Finnish, Kielipankki Version
17. The Letters of Paul Sinebrychoff, Kielipankki Version
18. The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
19. Classics Library of the National Library of Finland - Kielipankki version
20. Virtual Old Literary Finnish (VVKS) - Kielipankki Korp version
21. The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874)

18 corpora are PoS/MSD-tagged, 13 corpora are lemmatised and 6 corpora are syntactically parsed.

2.1.3.4. Licence

Licence information is available for 60 corpora. The following 8 corpora lack such information:

1. Hungarian Historical Corpus
2. Historical Corpus of the Welsh Language 1500-1850
3. "PolDiLemma" Middle Polish Diachrone Lemmatised Corpus
4. Mannheimer Korpus Historischer Zeitungen und Zeitschriften
5. Austrian Baroque Corpus
6. Hansard Corpus
7. Die Grenzboten (journal)
8. Brieven als buit (Letters as loot)

35 corpora are available under CC-BY, 11 corpora under the Oxford Text Archive licence, 6 under CLARIN ACA, 3 under EUPL v.1.1 SA, the rest have miscellaneous licences.

3. Non-CLARIN historical corpora

Table 2 lists 25 corpora that are not part of the CLARIN infrastructure.

Table 2: non-CLARIN historical corpora

Corpus	Language	Description
DIAKORP v6 **Size:** 4 million tokens **Annotation:** basic structural markup **Licence:** CC-BY-NC-SA	Czech	This corpus contains texts from the 14th to the 20th century. The corpus is available through a dedicated concordancer.
Brieven als buit (Letters as loot) **Size:** 460,000 words **Annotation:** lemmatised, PoS-tagged, grammatically tagged	Dutch	This corpus contains 1,000 letters from the 17th to the 18th century. The corpus is available through a dedicated concordancer.
ARCHER Corpus	English	The corpus contains texts from 1600 to 1999. Availability of the corpus is restricted.
ECCO-TCP **Size:** 74 million tokens **Annotation:** no linguistic annotation **Licence:** CC-1	English	This corpus contains texts (literature, philosophy, politics, religion, geography, science and all other areas of human endeavour) from 1700 to 1800. The corpus is available for download from the University of Oxford Text Archive and through a dedicated concordancer.
EEBO-TCP **Size:** 766 million tokens **Annotation:** no linguistic annotation **Licence:** CC-0	English	This corpus contains texts (literature, philosophy, politics, religion, geography, science and all other areas of human endeavour) from 1450 to 1750. The corpus is available for download from the University of Oxford Text Archive and through a dedicated concordancer.
EVANS-TCP **Size:** 766 million tokens **Annotation:** no linguistic annotation **Licence:** CC-0	English	This corpus contains American texts from 1640 to 1821. The corpus is available for download from the University of Oxford Text Archive and through a dedicated concordancer.
Historical Corpora at Lancaster University **Annotation:** tokenised, PoS-tagged, partial semantic tagging (USAS system)	English	The corpus contains texts in various domains (e.g., fiction, newspaper texts, religious texts) from 1500 on. The corpus is available through the CQPConcordancer.

Frantext **Size:** 300 million words **Annotation:** PoS-tagged, lemmatised	French	This corpus contains texts from the 10th to the 21st century. The corpus is available through a dedicated concordancer.
Bundesblatt/Feuille fédérale/Foglio federale **Size:** 203,585,806 tokens (German), 239,125,036 tokens (French), 85,223,085 tokens (Italian) **Annotation:** tokenised, syntactically-parsedd	German, French, Italian	This corpus contains texts from 1849 to 2014. The corpus is available through the CQPWeb concordancer.
Corpus of Old and Middle Hungarian court records and private correspondence **Size:** 850,000 words **Annotation:** tokenised, MSD-tagged, lemmatised, sociolinguistic metadata	Hungarian	This corpus contains private letters and testimonies from the 16th to the 18th century. The corpus is available through a dedicated concordancer.
Old Hungarian Corpus **Size:** 3 million tokens **Annotation:** tokenised, partially normalized, partially MSD-tagged	Hungarian	This corpus contains texts (codices, letters) from the 12th to the 17th century. The corpus is available for download from a dedicated webpage and through a dedicated concordancer.
Archivio Datini **Annotation:** lemmatised	Italian	This corpus contains Italian texts. The corpus is available through a dedicated concordancer.
Corpus testuale del Tesoro della Lingua Italiana delle Origini **Size:** 23 million tokens **Annotation:** tokenised, lemmatised	Italian	This corpus contains early Italian texts before 1375. The corpus is available through a dedicated concordancer.

DiaCORIS	Italian	This corpus contains texts from 1861 to 1945. The corpus is available through a dedicated concordancer.
M.I.DIA. (Morfologia dell'Italiano in DIAcronia) **Size:** 7.5 million tokens **Annotation:** tokenised **Licence:** CC-BY-NC 4.0	Italian	This corpus contains texts from the 13th to the 20th century. The corpus is available through a dedicated concordancer
Chronopress **Size:** 16 million tokens **Licence:** CC-BY-SA	Polish	This corpus contains newspaper articles from 1945 to 1954. The corpus is available through a dedicated concordancer.
Corpus of the 19. century Polish (Korpus polszczyzny XIX-wiecznej) **Size:** 625,000 tokens **Annotation:** tokenised, PoS-tagged, lemmatised, transliteration, transcripton	Polish	This corpus contains texts from 1830 to 1918. The corpus is available through a dedicated concordancer.
IMPACT GT corpus (Korpus GT projektu IMPACT) **Size:** 1.5 million tokens **Annotation:** transcription	Polish	This corpus contains texts from 1570 to 1756. The corpus is available through a dedicated concordancer.
Corpus of old Polish texts until 1500 (Korpus tekstów staropolskich do roku 1500) **Size:** 620,000 tokens **Annotation:** tokenised	Polish, Latin	This corpus contains texts until 1500. The corpus is available for download from a dedicated webpage.
Corpus of the 16. century Polish (Korpus polszczyzny XVI wieku) **Annotation:** lemmatised, transliteration	Polish, Latin	This corpus contains texts from the 16th century. The corpus is available through a dedicated concordancer.
eFontes Mediae et Infimae Latinitatis Polonorum (Elektroniczny korpus polskiej łaciny średniowiecznej) **Size:** 5 million tokens **Annotation:** tokenised, lemmatised	Polish, Latin	This corpus contains texts from the 11th to the middle of the 16th century. The corpus is available through a dedicated concordancer.

The Electronic Corpus of the 17th and 18th century Polish (Korpus tekstów polskich z XVII i XVIII w.) **Size:** 12 million tokens **Annotation:** tokenised, partially PoS-tagged, structural annotation	Polish, Latin	This corpus contains texts from 1601 to 1772. The corpus is available through a dedicated concordancer.
XV century New Testament translations (Piętnastowieczne przekłady Nowego Testamentu – elektroniczna konkordancja staropolska) **Size:** 400,000 tokens **Annotation:** tokenised	Polish, Latin	This corpus contains Biblical texts from 1380 to 1500. This corpus is available for download from a dedicated webpage and through a dedicated concordancer.
Corpus Informatizado do Português Medieval **Size:** 2 million tokens **Annotation:** tokenised, PoS-tagged	Portuguese	This corpus contains texts from the 9th to the 16th century. The corpus is available through a dedicated concordancer.
Parsed Corpus of Historical Portuguese **Size:** 3.3 million **Annotation:** tokenised, PoS-tagged (2 million), treebanked (1.2 million)	Portuguese	This corpus contains 76 texts written by authors born between 1380 and 1881. The corpus is available through a dedicated concordancer.

3.1. Summary

3.1.1. Availability

The following 5 corpora are available for download and through a concordancer:

1. XV century New Testament translations (Piętnastowieczne przekłady Nowego Testamentu – elektroniczna konkordancja staropolska)
2. Old Hungarian Corpus
3. ECCO-TCP
4. EEBO-TCP
5. EVANS-TCP

The following 18 corpora are available through a concordancer.

1. Corpus Informatizado do Português Medieval
2. Parsed Corpus of Historical Portuguese
3. Historical Corpora at Lancaster University
4. Corpus testuale del Tesoro della Lingua Italiana delle Origini
5. DiaCORIS
6. M.I.DIA. (Morfologia dell'Italiano in DIAcronia)
7. Archivio Datini
8. Frantext
9. eFontes Mediae et Infimae Latinitatis Polonorum (Elektroniczny korpus polskiej łaciny średniowiecznej)
10. Corpus of the 16. century Polish (Korpus polszczyzny XVI wieku)
11. The Electronic Corpus of the 17th and 18th century Polish (Korpus tekstów polskich z XVII i XVIII w.)
12. Corpus of the 19. century Polish (Korpus polszczyzny XIX-wiecznej)
13. IMPACT GT corpus (Korpus GT projektu IMPACT)
14. Chronopress
15. Bundesblatt/Feuille fédérale/Foglio federale
16. DIAKORP v6
17. Corpus of Old and Middle Hungarian court records and private correspondence
18. Brieven als buit (Letters as loot)

One corpus is available for download:

1. Corpus of old Polish texts until 1500 (Korpus tekstów staropolskich do roku 1500)

3.1.2. Metadata

3.1.2.1. Language

There are 19 monolingual corpora in Table 2 for the 8 following languages:

1. English (5 corpora)
2. Italian (4 corpora)
3. Polish (3 corpora)
4. Hungarian (2 corpora)
5. Portuguese (2 corpora)
6. Czech (1 corpus)
7. Dutch (1 corpus)
8. French (1 corpus)

There are 6 multilingual corpora for the following 2 language combinations:

1. Polish, Latin (5 corpora)
2. German, French, Italian (1 corpus)

3.1.2.2. Size and temporal span

Information on size is available for the majority of the corpora – 17 out of 22. The following corpora 5 corpora lack this information:

1. ARCHER Corpus
2. Historical Corpora at Lancaster University
3. DiaCORIS
4. Archivio Datini
5. Corpus of the 16. century Polish (Korpus polszczyzny XVI wieku)

The oldest corpus reaches back to the 9th century.

3.1.2.3. Annotation

Information on annotation is likewise available for the majority of the corpora – 19 out of 22. The following 3 corpora lack this information:

1. ARCHER Corpus
2. DiaCORIS
3. Corpus of the 19. century Polish (Korpus polszczyzny XIX-wiecznej)

3.1.2.4. Licence

Licence information is available for 4 corpora:

1. Chronopress
2. ECCO-TCP
3. EEBO-TCP
4. EVANS-TCP

4. Linguistic tools for the annotation of historical corpora

The following tools/services were contributed by a NC/UI coordinator in the Google Spreadsheet. All are available through a CLARIN-D node:

1. Part-of-speech tagging: mixed text
   Description: A webapplication that reads in mixed texts written in Latin and Middle English in txt-Format
2. TreeTagger -- Middle High German parameter file
   Description: A trained parameter file for the TreeTagger (part-of-speech tagger) to tag Middle High German text.
3. OCR Post-Correction
   Description: A web application that reads in scans of German books from the 18th to 20th century in Antiqua and Fraktur and runs a pipeline of OCR with tesseract and a post-correction system using different techniques.