Title	Report on the CLARIN Resource Families, Volume 2
Version	2.0
Author(s)	DF, JL
Date	9-11-2020
Status	For distribution
Distribution	BoD, NCF, UI, SAB, GA
ID	CE-2018-1236

Contents

1. Introduction

This report provides detailed summaries and an in-depth overview of the issues identified with surveys of the CLARIN ERIC Resource Families available in the VLO and the CLARIN repositories, such as findability, accessibility and missing metadata.

Since the Resource Families were launched in April 2018, they have become one of the flagship initiatives of User Involvement in CLARIN ERIC. The Resource Families now comprise 12 corpus families (parliamentary, computer-mediated communication, parallel, newspaper, L2-learner, historical, spoken, manually curated, literary, academic, reference, and multimodal corpora), 5 families of lexical resources (lexica, dictionaries, conceptual resources, glossaries, and wordlists), and 4 families of language tools (tools for normalization, tools for named entity recognition, part-of-speech taggers and lemmatizers, and tools for sentiment analysis), which together amount to 974 manually curated tools and resources as of 5 November 2020. The individual resource and tool families are generally also top-ranked Google results for searches that include associated keywords.[1]

This report is an update of the Report on the CLARIN corpora in CLARIN Resource Families published in June 2018 which focused on the first 6 corpus families, and of the paper published in the proceedings of the 2020 CLARIN Annual Conference (Lenardič and Fišer 2020), in which we present an earlier version that reflected the state of this initiative current on 17 August 2020. At that time, the initiative did not yet include the multimodal corpora family. Several updates have also been made on the other families of resources and tools. This updated report extends the overview for all subsequently surveyed resources and tools as well as provides a detailed changelog for all the issues that have been resolved since the publication of the first report.

The report is organized as follows. In Section 2, we provide summaries, issue listings, and changelogs for all the 12 corpus families. In Section 3, we do the same for the 5 lexical resource families. Section 4 does the same for the 4 tool families. Section 5 provides a detailed summary. Section 6 concludes the report.

[1] Check, for instance, the search terms annotated corpus and parliamentary corpus in Google.

2. Corpora

2.1 Parliamentary corpora

2.1.1 Summary

In Table 1, we summarize the information on parliamentary corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, annotation and licence. The summary is based on the Parliamentary corpora subpage of the Resource Families that was last updated on 30 October 2020.

Table 1: Summary of information on parliamentary corpora within the CLARIN infrastructure

Identification	• 25 parliamentary corpora part of the CLARIN infrastructure in total • 22 (88%) corpora identified through the VLO • 3 (12%) corpora identified through national repositories, but not through VLO
Availability	• 10 (40%) corpora for download and through a concordancer • 11 (44%) corpora for download • 3 (12%) corpora through a concordancer • 1 (4%) corpus unavailable
Languages	• 22 (91%) corpora are monolingual, 3 (9%) are multilingual • 3 Slovenian, 2 English, 2 German, 2 Greek, 2 Norwegian corpora • 1 corpus per language: Croatian, Czech, Danish, Estonian, Finnish, French, Icelandic, Lithuanian, Polish, Portuguese, Swedish
Size	• Information on size is available for all corpora • Largest corpus has 1.6 billion tokens, smallest has 190,000 tokens • 7 small corpora (<10 million words/tokens) • 12 medium-sized corpora (10–100 million words/tokens) • 6 large corpora (>100 million words/tokens)
Annotation	• 12 (48%) corpora PoS/MSD-tagged, 12 (48%) corpora lemmatised • Unknown for 3 (12%) corpora
Licence	• 16 (67%) corpora under CC-BY • Unknown for 3 (12%) corpora

2.1.2 Issues

In the following Table, we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 2: Issues

List of parliamentary corpora with metadata issues	List of CLARIN parliamentary corpora not in the VLO
1. Hansard corpus a. Missing licence info 2. ParlAT beta a. Missing licence info (corpus under construction) b. Unavailable due to being under construction 3. Hellenic Parliament Minutes (1989-1994, 1997-2018) a. Missing annotation info 4. Speeches of Politicians in the Greek Parliament a. Missing annotation info	1. Hellenic Parliament Minutes (1989-1994, 1997-2018) 2. Speeches of Politicians in the Greek Parliament 3. European Parliament Proceedings Parallel Corpus 1996-2011, parallel corpus Greek-English

5. Polish Parliamentary Corpus a. Missing licence info 6. Transcripts of Riigikogu (Estonian Parliament) a. Missing annotation info	

2.1.3 Changelog

The following issues have been resolved and changes made for this resource family since the last report on CLARIN corpora was published on 13 June 2018:

- 9 new parliamentary corpora have been added:
    1. Croatian parliamentary corpus ParlaMeter-hr 1.0
    2. The Danish Parliament Corpus 2009 - 2017, v1
    3. Hellenic Parliament Minutes (1989-1994, 1997-2018)
    4. Speeches of Politicians in the Greek Parliament, European Parliament Proceedings Parallel Corpus 1996-2011, parallel corpus Greek-English
    5. The Icelandic Parliamentary Corpus
    6. Slovenian parliamentary corpus siParl 1.0
    7. Slovenian parliamentary corpus ParlaMeter-sl 1.0
    8. Multilingual comparable corpora of parliamentary debates ParlaMint 1.0
- 1 additional corpus is now findable through the VLO: The Icelandic Parliamentary Corpus
- 3 corpora have been removed from the overview:
    1. Hungarian National Corpus, as it is not a parliamentary corpus
    2. Hellenic Parliamentary Sittings, as it is replaced by Hellenic Parliament Minutes (1989-1994, 1997-2018)
    3. DK-CLARIN Almensprogligt korpus – tekster fra Folketinget, as it is replaced by The Danish Parliament Corpus 2009 - 2017, v1
- Annotation info has been added for 1 corpus:
    1. Plenary Sessions of the Parliament of Finland (MSD-tagged, lemmatized, syntactically parsed)
- Licence info has been added for 2 corpora
    1. Europarl: European Parliament Proceedings Parallel Corpus 1996-2011 (CC0)
    2. Lithuanian Parliament Corpus for Authorship Attribution (CLARIN PUB)
- Period has been resolved for 1 corpus:
    1. Czech Parliamentary Meetings (15 November 2016–21 November 2018)

2.2. CMC corpora

2.2.1. Summary

In Table 3, we summarize the information on CMC corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, annotation and licence. The summary is based on the CMC corpora subpage of the Resource Families that was last updated on 26 October 2020.

Table 3: Summary of information on CMC corpora within the CLARIN infrastructure

Identification	• 15 CMC corpora part of the CLARIN infrastructure in total • 12 (80%) corpora identified through the VLO • 3 (20%) corpora identified through national repositories, but not through VLO
Availability	• 8 (53%) corpora for download and through a concordancer • 2 (14%) corpora through a concordancer • 5 (33%) corpora for download
Languages	• 14 (93%) corpora are monolingual, 1 (7%) is multilingual • 5 Slovenian corpora, 2 French corpora • 1 corpus per language: Czech, Dutch, Estonian, Finnish, German, Lithuanian, Norwegian
Size	• Information on size available for all corpora • Largest corpus has 2.6 billion tokens, smallest has 600,000 tokens • 5 small corpora (<10 million words/tokens) • 6 medium-sized corpora (10–100 million words/tokens) • 4 large corpora (>100 million words/tokens)
Annotation	• 9 (60%) corpora PoS/MSD-tagged, 8 (53%) corpora lemmatised • Unknown for 4 (27%) corpora
Licence	• 8 (53%) corpora under CC-BY • Unknown for 2 (13%) corpora

2.2.2. Issues

In the following Table, we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 4: Issues

List of CMC corpora with metadata issues	List of CLARIN CMC corpora not in the VLO
1. SoNaR New Media a. Missing licence info 2. The Mixed Corpus: New Media a. Missing annotation info b. Missing licence info 3. LITIS v.1 a. Missing annotation info 4. NTAP French a. Missing annotation info b. Missing licence info 5. NTAP English a. Missing annotation info b. Missing licence info	1. The Mixed Corpus: New Media

2.2.3. Changelog

The following issues have been resolved and changes made for this resource family since the last report on CLARIN corpora was published on 13 June 2018:

- 2 new corpora were added: NTAP French and NTAP English
- Annotation info has been added for the following 2 corpora:
    1. Corpus of contemporary blogs (sentence tagging, as described here)
    2. CoMeRe repository (no linguistic annotation)

2.3. Parallel corpora

2.3.1. Summary

In Table 5, we summarize the information on parallel corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, alignment and licence. The summary is based on the parallel corpora subpage of the Resource Families that was last updated on 30 October 2020.

Table 5: Summary of information on parallel corpora within the CLARIN infrastructure

Identification	• 86 parallel corpora part of the CLARIN infrastructure in total • 53 (62%) corpora identified through the VLO • 33 (38%) corpora identified through national repositories, but not through VLO
Availability	• 5 (6%) corpora for download and through a concordancer • 63 (73%) corpora for download • 8 (9%) corpora through a concordancer • 10 (12%) corpora unavailable
Languages	• 47 (55%) corpora are bilingual, 39 (45%) are multilingual • Biggest number of languages in a single corpus: 117 • 6 corpora for more than 50 languages
Size	• Missing for 11 (13%) corpora • Largest corpus has 4.2 billion tokens, smallest has 43,000 tokens • 29 small corpora (<10 million words/tokens) • 16 medium-sized corpora (10–100 million words/tokens) • 8 large corpora (>100 million words/tokens)
Alignment	• 52 (60%) corpora sentence-aligned; 5 (6%) corpora word-aligned • Unknown for 27 (31%) corpora
Licence	• 37 (43%) corpora under CC-BY • Missing for 6 (7%) corpora

2.3.2. Issues

In the following Table, we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 6: Issues

List of parallel corpora with metadata issues	List of CLARIN parallel corpora not in the VLO
1. Amharic-English bilingual corpus a. Missing alignment info 2. Czech-Slovak Parallel Corpus a. Missing alignment info 3. Kacenka a. Missing alignment info b. Missing licence info c. Unavailable 4. QTLP English-Greek Corpus for the MEDICAL domain a. Unavailable despite download option 5. QTLP English-Greek Corpus for the AUTOMOTIVE domain a. Unavailable despite download option	1. Parallel corpus newsletters IFT FR-GR 2. ACCURAT balanced test corpus for under resourced languages 3. European Parliament Proceedings Parallel Corpus 1996-2011, parallel corpus Greek-English 4. EMEA Corpus 5. ECDC Translation Memory 6. DGT-Translation Memory 7. DGT-Acquis 8. EAC Translation Memory 9. A parallel corpus collected from the European Constitution 10. A parallel corpus of KDE4 localization files (v.2) 11. European Central Bank parallel corpus

6. Inlerlingual Perspectives[2]
    a. Missing alignment info
7. Text Corpus - EMEL
    a. Missing alignment info
8. aformes
    a. Missing alignment info
9. The English-Slovak Parallel corpus
    a. Missing size info
    b. Missing alignment info
10. English-Luganda Parallel Corpus
    a. Missing licence info
11. Parallel corpus newsletters IFT FR-GR
    a. Missing size info
    b. Missing alignment info
12. FREL
    a. Missing alignment info
    b. Unavailable despite download option
13. QTLP German-Greek Corpus for the MEDICAL domain
    a. Unavailable despite download option/licence
14. LOGON parallel tourist corpus of Norwegian-English texts
    a. Missing alignment info
    b. Missing licence info
15. The Norwegian-Spanish Parallel Corpus
    a. Missing alignment info
16. The Polish-Lithuanian Parallel Corpus
    a. Missing size info
    b. Missing alignment info
17. COMPARA : Portuguese - English parallel translation corpus
    a. Missing size info
18. QTLP Portuguese-Greek Corpus for the MEDICAL domain
    a. Unavailable despite download option/licence
19. QTLP Portuguese-Greek Corpus for the AUTOMOTIVE domain
    a. Unavailable despite download option/licence
20. Parallel Bible Corpus
    a. Missing size info
    b. Missing alignment info
    c. Missing licence info
    d. Unavailable due to broken and incorrect link
21. DGT-Acquis

12. OpenSubtitles2011
13. SPC - Stockholm Parallel Corpora
14. Tatoeba
15. DGT-TM-2016
16. QTLP English-Greek Corpus for the MEDICAL domain
17. QTLP German-Greek Corpus for the MEDICAL domain
18. QTLP Portuguese-Greek Corpus for the MEDICAL domain
19. QTLP English-Greek Corpus for the AUTOMOTIVE domain
20. QTLP Portuguese-Greek Corpus for the AUTOMOTIVE domain
21. Text Corpus - EMEL
22. FREL
23. Inlerlingual Perspectives
24. aformes
25. GLOSSOLOGIA
26. Civitas Gentium
27. Official Journal of the European Union
28. INTERA Corpus - the Greek-English part
29. Greek-Bulgarian Bul-TM parallel corpus
30. LOGON parallel tourist corpus of Norwegian-English texts
31. PELCRA Polish-English parallel corpora
32. The Corpus of Free Trade Agreement
33. UP/TAP annotated by the OpenNLP Part-of-Speech Tagger (Portuguese) and OpenNLP Part-of-Speech Tagger (English)

[2] Note that the original CLARIN:EL entry of the resource is misspelt.

a. Missing size info 22. Official Journal of the European Union a. Missing size info b. Unavailable despite download option c. Unclear “other” licence 23. DGT-Translation Memory a. Missing alignment info 24. ParaCrawl Corpus version 1.0 a. Missing size info b. Missing alignment info 25. MLCC Multilingual and Parallel Corpora a. Missing alignment info 26. The CLUVI parallel corpus a. Missing alignment info 27. Europarl QTLeap WSD/NED corpus a. Missing alignment info 28. GLOSSOLOGIA a. Missing size info b. Missing alignment info 29. MULCOLD - Multilingual Corpus of Legal Documents a. Missing alignment info 30. CRATER 2 Corpus a. Missing alignment info 31. CsEnVi Pairwise Parallel Corpora a. Missing alignment info 32. The DPC – Dutch Parallel Corpus a. Missing licence info 33. EuroParl-UdS a. Missing size info 34. PANACEA English-French and English-Greek parallel corpus a. Missing size info b. Missing alignment info 35. Polish-Bulgarian-Russian Parallel Corpus a. Missing alignment info b. Missing licence info 36. MUSA Multilingual Multimodal Corpus a. Unavailable despite download option 37. The Corpus of Free Trade Agreement a. Missing alignment info 38. European Parliament Interpretation Corpus (EPIC) a. Missing alignment info 39. Civitas Gentium a. Missing alignment info	

2.3.3. Changelog

The following issues have been resolved and changes made for this resource family since the last report on CLARIN corpora was published on 13 June 2018:

- The following 5 corpora have been added:
    1. UP/TAP annotated by the OpenNLP Part-of-Speech Tagger (Portuguese) and OpenNLP Part-of-Speech Tagger (English)
    2. ParIce
    3. ParaCrawl Corpus version 1.0
    4. MultiJur: Multilingual Parallel Corpus of Legal Texts
    5. EuroParl-UdS
- The following corpus can now be found through the VLO:
    1. Opus, Helsinki Korp Version
- Size info has been added for 3 corpora:
    1. Polish-Bulgarian-Russian Parallel Corpus (55 texts)
    2. English-Urdu Religious Parallel Corpus (14,371 sentences)
    3. EMEA Corpus (31 million tokens)
- Alignment info has been added for 4 corpora:
    1. HindEnCorp 0.5 (sentence-aligned)
    2. English-Czech Corpus from Wikipedia (sentence-aligned)
    3. Serbian-English parallel corpus srenWaC 1.0 (sentence-aligned)
    4. Tourism English-Croatian Parallel Corpus (sentence-aligned)
- Licence was resolved for 1 corpus:
    - JRC-Acquis Multilingual Parallel Corpus (Usage/licencing conditions, as described here)

2.4. Newspaper corpora

2.4.1. Summary

In Table 7, we summarize the information on newspaper corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, annotation and licence. The summary is based on the newspaper corpora subpage of the Resource Families that was last updated on 26 October 2020.

Table 7: Summary of information on newspaper corpora within the CLARIN infrastructure

Identification	• 33 newspaper corpora part of the CLARIN infrastructure in total • 22 (67%) corpora identified through the VLO • 11 (33%) corpora identified through national repositories, but not through VLO
Availability	• 11 (33%) corpora for download and through a concordancer • 16 (48%) corpora for download • 5 (15%) corpora through a concordancer • 1 (3%) corpora unavailable
Languages	• 26 (79%) corpora are monolingual, 7 (21%) are multilingual • 11 Swedish, 4 German, 3 Greek, 2 Czech, 2 French Corpora • 1 corpus per language: Arabic, Finnish, Norwegian, Polish
Size	• Information on size missing for 2 (6%) corpora • Largest corpus has 8.8 billion tokens, smallest has 500,000 tokens • 17 small corpora (<10 million words/tokens) • 3 medium-sized corpora (10–100 million words/tokens) • 8 large corpora (>100 million words/tokens)
Annotation	• 18 (55%) corpora PoS/MSD-tagged, 12 (36%) corpora syntactically parsed • Unknown for 14 (42%) corpora
Licence	• 19 (58%) corpora under CC-BY • Unknown for 3 (9%) corpora

2.4.2. Issues

In the following Table, we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 8: Issues

List of newspaper corpora with metadata issues	List of CLARIN newspaper corpora not in the VLO
1. An-Nahar Newspaper Text Corpus a. Missing annotation info 2. The Karelian Finnish Newspaper Corpus a. Missing annotation info 3. BREF-80 a. Missing annotation info 4. Corpus journalistique issu de l'Est Républicain a. Missing size info 5. Mannheim Corpus of Historical Newspapers and Magazines a. Missing annotation info b. Missing licence info 6. Corpus "Library and Information Centre - Newspapers" a. Missing annotation info	1. DN 1987 2. GP 1994 and 2001-2011 3. Kubhist 4. The Webbnyheter corpus 5. Dagny 6. Hertha 7. Idun 8. Kvinnorans Tidning 9. Morgonbris 10. Smittskydd 11. Rösträtt för Kvinnor

b. Missing licence info 7. Modern Greek Texts Corpus - "Makedonia" newspaper a. Missing annotation info 8. Modern Greek Texts Corpus - "Ta Nea" newspaper a. Missing annotation info 9. The Norwegian Newspaper Corpus a. Missing annotation info (only specified as “multitagged”) b. Missing licence info 10. Parallel Global Voices a. Missing annotation info 11. MLCC Multilingual and Parallel Corpora a. Missing annotation info 12. ACCURAT corpus of comparable sentences a. Missing annotation info 13. The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version a. Missing annotation info 14. The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874) a. Missing size info b. Missing annotation info 15. Corpora of Newspaper Texts a. Unavailable b. Missing annotation info	

2.4.3. Changelog

The following issues have been resolved and changes made for this resource family since the last report on CLARIN corpora was published on 13 June 2018:

- Size info has been added for 2 corpora:
  1. The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version (8.8 billion tokens)
  2. An-Nahar Newspaper Text Corpus (24 million tokens)
- Annotation info has been added for the following corpus:
  1. Corpus journalistique issu de l'Est Républicain (MSD-tagged, lemmatised)

2.5. L2-learner corpora

2.5.1. Summary

In Table 9, we summarize the information on L2-learner corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, annotation and licence. The summary is based on the L2-learner corpora subpage of the Resource Families that was last updated on 30 October 2020.

Table 9: Summary of information on L2-learner corpora within the CLARIN infrastructure

Identification	• 74 L2 learner corpora part of the CLARIN infrastructure in total • 72 (98%) corpora identified through the VLO • 2 (2%) corpora identified through national repositories, but not through VLO
Availability	• 41 (55%) corpora for download and through a concordancer • 18 (25%) corpora for download • 6 (8%) corpora through a concordancer • 9 (12%) corpora unavailable
Languages	• 63 (85%) corpora are monolingual, 11 (15%) are multilingual • 20 English, 12 Spanish, 10 French, 5 German, 4 Finnish, 3 Swedish, 2 Czech, 2 Mandarin corpora • 1 corpus per language: Arabic, Hungarian, Icelandic, Italian, Norwegian
Size	• Information on size missing for 8 (11%) corpora • Largest corpus has 3 million tokens, smallest has 24,000 tokens • 50 small corpora (<1 million words/tokens) • 4 medium-sized corpora (≥1 million words/tokens)
Annotation	• 36 (49%) corpora with audio/transcription linking, 10 (14%) corpora PoS/MSD-tagged, 8 (11%) corpora error mark-up • Unknown for 20 (27%) corpora
Licence	• 50 (67%) corpora under CC-BY, 12 (16%) CLARIN RES • Unknown for 4 (5%) corpora

2.5.2. Issues

In the following Table, we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 10: Issues

List of L2 learner corpora with metadata issues	List of CLARIN L2 learner corpora not in the VLO
1. British Academic Written English Corpus a. Missing annotation info 2. ETS Corpus of Non-Native Written English a. Missing annotation info 3. The Hanken Corpus of Academic Writing a. Missing annotation info b. Unavailable 4. ICLE International Corpus of Learner English a. Missing annotation info b. Unavailable c. Missing licence info 5. The Uppsala Student English corpus a. Missing annotation info 6. Testipiste Corpus	1. SW1203-essays 2. Tisus corpus

    a. Missing annotation info
    b. Unavailable

7. Commented Learner Corpus Academic Writing
    a. Missing annotation info
8. The Anglish Corpus
    a. Missing size info
9. GLBCC (Giessen - Long Beach Chaplin Corpus)
    a. Missing annotation info
10. A Learners' Corpus of Reading Texts
    a. Missing size info
    b. Missing annotation info
11. Arabic Learner Corpus
    a. Missing annotation info
12. English as a Foreign Language Corpus
    a. Missing annotation info
    b. Unavailable
13. The Long Second Corpus
    a. Missing size info
    b. Missing annotation info
    c. Unavailable
14. CEFLING Project Corpus
    a. Missing size info
    b. Missing annotation info
    c. Missing licence info
    d. Unavailable (“page-not-found”)
15. DIALUKI: Diagnosing reading and writing in a second or foreign language
    a. Missing annotation info
    b. Unavailable
16. Topling - Paths in Second Language Acquisition
    a. Missing annotation info
17. LeaP: The Learning the Prosody of a Foreign Language
    a. Missing licence info
18. Repiso/Contrefactualité
    a. Missing size info
    b. Missing annotation info
19. Openprodat
    a. Missing size info
    b. Missing annotation info
20. GeWiss
    a. Missing licence info
21. TAITO: Written and Oral Data of the TAITO-project
    a. Missing size info
    b. Missing annotation info
    c. Unavailable
22. YKI National Certificates corpus

a. Missing size info b. Missing annotation info c. Unavailable 23. AixOx a. Missing annotation info	

2.5.3. Changelog

The following issues have been resolved and changes made for this resource family since the last report on CLARIN corpora was published on 13 June 2018:

- In early 2020, 38 TalkBank corpora were temporarily no longer listed in the VLO; however, this issue was remedied in September 2020.
- Annotation info has been added for the following 3 corpora:
    1. LANGMAN (error coding)
    2. FinSveStud 79-80 (lemmatised)
    3. The Advanced Finnish Learners' Corpus (MSD-tagged, lemmatised)
- Licence info has been added for the following 2 corpora:
    1. LANGMAN (CC-BY)
    2. The Anglish Corpus (CLARIN RES)

2.6. Historical corpora

2.6.1. Summary

In Table 11, we summarize the information on historical corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, annotation and licence. The summary is based on the historical corpora subpage of the Resource Families that was last updated on 30 October 2020.

Table 11: Summary of information on historical corpora within the CLARIN infrastructure

Identification	• 74 historical corpora part of the CLARIN infrastructure in total • 68 (92%) corpora identified through the VLO • 6 (8%) corpora identified through national repositories, but not through VLO
Availability	• 15 (20%) corpora for download and through a concordancer • 32 (43%) corpora for download • 24 (32%) corpora through a concordancer • 3 (5%) corpora unavailable
Languages	• 57 (77%) corpora are monolingual, 17 (23%) are multilingual • 17 English, 13 German, 6 Finnish, 2 Dutch, 2 French, 2 Polish, 2 Greek, 2 Scots, 2 Slovenian • 1 corpus per language: Akkadian, Chinese, Hungarian, Icelandic, Latin, Norse, Swedish, Welsh
Size	• Missing for 12 (16%) corpora • Largest corpus has 8.7 billion tokens, smallest has 4000 tokens • 42 small corpora (<10 million words/tokens) • 9 medium-sized corpora (10–100 million words/tokens) • 8 large corpora (>100 million words/tokens)
Annotation	• 24 (32%) corpora PoS/MSD-tagged, 17 (23%) corpora lemmatized • Unknown for 18 (24%) corpora
Licence	• 43 (58%) corpora under CC-BY • Missing for 6 (8%) corpora

2.6.2. Issues

In the following Table, we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 12: Issues

List of historical corpora with metadata issues	List of CLARIN historical corpora not in the VLO
1. Greek Medieval Texts a. Missing annotation info (under-specified as “other”) 2. Sheffield Corpus of Chinese a. Missing size info 3. Corpus of Late Modern English prose / David Denison a. Missing size info 4. Hansard Corpus a. Missing licence info 5. Pamphlets of the American Revolution : [selections] / edited by Bernard Bailyn a. Missing size info 6. The Lampeter Corpus of Early Modern English Tracts	1. Greek Medieval Texts 2. Austrian Baroque Corpus 3. OROSSIMO Corpus - History 4. DDR-Presseportal (GDR press portal) 5. Classics Library of the National Library of Finland - Kielipankki version 6. Virtual Old Literary Finnish (VVKS) - Kielipankki Korp version

   a. Missing size info

7. The Lancaster Newsbooks Corpus
   a. Missing size info
8. Classics of Finnish Literature, Kielipankki Version
   a. Missing annotation info
9. Corpus of Old Literary Finnish
   a. Missing annotation info
10. The Finnish Gutenberg Corpus
    a. Missing annotation info
11. Virtual Old Literary Finnish (VVKS) - Kielipankki Korp version
    a. Missing annotation info
    b. Unavailable
12. Partonopeus de Blois: transcriptions of all manuscripts and fragments
    a. Missing size
13. Austrian Baroque Corpus
    a. Missing licence info
14. DDR-Presseportal (GDR press portal)
    a. Missing size info
    b. Missing annotation info
    c. Missing licence info
15. Mannheimer Korpus Historischer Zeitungen und Zeitschriften
    a. Missing annotation info
    b. Missing licence info
16. Hungarian Historical Corpus
    a. Missing annotation info
    b. Missing licence info
17. Chronopress
    a. Missing annotation info
18. Corpus of biblical text in Scots / John Kirk
    a. Missing annotation info
19. The Helsinki corpus of Older Scots : [1450-1700]
    a. Missing size info
20. Historical Corpus of the Welsh Language 1500-1850
    a. Missing annotation info
    b. Missing licence info
    c. Unavailable (decommissioned service)
21. Dictionary of Old English Corpus in Electronic Form (DOEC)
    a. Missing size info
22. The Electronic Text Corpus of Sumerian Literature. Revised edition.
    a. Missing size info
23. Finnish Folk Poetry

a. Missing annotation info 24. Corpus of Early Modern Finnish, Kielipankki Version a. Missing annotation info 25. Aleksis Kivi Corpus (SKS) a. Missing annotation info 26. Classics Library of the National Library of Finland - Kielipankki version a. Missing size info b. Missing annotation info c. Unavailable (not released yet) 27. The Letters of Paul Sinebrychoff, Kielipankki Version a. Missing annotation info 28. The Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version a. Missing annotation info 29. The Newspaper and Periodical OCR Corpus of the National Library of Finland (1771-1874) a. Missing size info b. Missing annotation info	

2.6.3. Changelog

The following issues have been resolved and changes made for this resource family since the last report on CLARIN corpora was published on 30 June 2018:

- 2 new corpora were added: The Saga Corpus and LatinISE corpus (version 4)
- 4 corpora that originally did not have VLO entries now have them:
    1. The Swedish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
    2. The Finnish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version
    3. The Morpho-Syntactic Database of Mikael Agricola's Works
    4. The Saga Corpus
- Size info has been added for the following 3 corpora:
    1. Virtual Old Literary Finnish (VVKS) (48 texts)
    2. Corpus of biblical text in Scots / John Kirk (35,506 words)
    3. “PolDiLemma” Middle Polish Diachrone Lemmatized Corpus (7 million words)
- Annotation info has been added for the following 13 corpora:
    1. A Corpus of English Dialogues (no annotation)
    2. Dictionary of Old English Corpus in Electronic Form (DOEC) (no annotation)
    3. Partonopeus de Blois: transcriptions of all manuscripts and fragments (no annotation)
    4. The Lampeter Corpus of Early Modern English Tracts (no annotation)
    5. The Lancaster Newsbooks Corpus (no annotation)
    6. Deutsches Textarchiv (PoS-tagged, lemmatised)
    7. Sheffield Corpus of Chinese (no annotation)
    8. Pamphlets of the American Revolution : [selections] / edited by Bernard Bailyn (no annotation)

9. The Helsinki corpus of Older Scots : [1450-1700] (no annotation)
10. Corpus of biblical text in Scots / John Kirk (no annotation)
11. GerManC. A Historical Corpus of German Newspapers 1650 (no annotation)
12. Anthology of Middle English texts / Santiago Gonzalez y Fernandez-Corugedo (no annotation)
13. Helsinki corpus of English texts (no annotation)

- Licence info has been added for the following 3 corpora:
    1. Brieven als buit (Letters as loot) (CLARIN PUB)
    2. Die Grenzboten (journal) (CC-BY-NC-SA 3.0)
    3. Deutsches Textarchiv (CLARIN PUB)

2.7. Spoken corpora

2.7.1. Summary

In Table 13, we summarize the information on spoken corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, annotation and licence. The summary is based on the spoken corpora subpage of the Resource Families that was current on 26 October 2020.

Table 13: Summary of information on spoken corpora within the CLARIN infrastructure

Identification	• 90 spoken corpora part of the CLARIN infrastructure in total • 89 (99%) corpora identified through the VLO • 1 (1%) corpora identified through national repositories, but not through VLO
Availability	• 35 (38%) corpora for download and through a concordancer • 31 (34%) corpora for download • 15 (17%) corpora through a concordancer • 9 (10%) corpora unavailable
Languages	• 71 (79%) corpora are monolingual, 19 (21%) are multilingual • 18 German, 9 Czech, 8 Estonian, 7 Finnish, 5 French, 5 Norwegian, 4 English, 2 Dutch, 2 Hungarian, 2 Slovenian, 2 Spanish corpora • 1 corpus per language: Arabic, Icelandic, Italian, Nepali, Polish, Saami, Swedish
Size	• Information on size missing for 12 (13%) corpora • Largest corpus in terms of audio length has 12,500 hours of recordings, smallest 2 hours • 8 small corpora (<10 hours of recordings) • 26 medium-sized corpora (10–100 hours of recordings) • 13 large corpora (>100 hours of recordings)
Annotation	• 37 (42%) corpora PoS/MSD-tagged, 37 (42%) corpora orthographically transcribed, 12 (13%) corpora phonetically transcribed • Unknown for 26 (29%) corpora
Licence	• 28 (31%) corpora under CLARIN RES, 25 (28%) under CC-BY • Unknown for 5 (6%) corpora

2.7.2. Issues

In the following Table, we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 14: Issues

List of spoken corpora with metadata issues	List of CLARIN spoken corpora not in the VLO
1. Arabic Speech Corpus a. Missing size info b. Missing annotation info 2. Spoken corpus of Karel Makoň a. Missing annotation info 3. Buckeye Corpus of Conversational Speech a. Missing size info 4. ELFA Corpus a. Missing annotation info 5. Corpus of Radio News a. Missing licence info 6. Estonian Emotional Speech Corpus a. Missing annotation info	1. Gothenburg Dialogue Corpus

7. Estonian North Wind and the Sun Corpus v.1.0.3 a. Missing size info b. Missing annotation info 8. Faroese Danish Corpus Hamburg 0.2.dan (FADAC-0.2.dan Hamburg) a. Missing size info 9. Aalto University DSP Course Conversation Corpus 2013-2016, Downloadable Version a. Missing annotation info 10. Finnish Broadcast Corpus a. Missing annotation info 11. Samples of Spoken Finnish a. Missing size info b. Missing annotation info 12. The Longitudinal Corpus of Finnish Spoken in Helsinki (1970s, 1990s and 2010s) a. Missing annotation info 13. The Corpus of Border Karelia a. Missing annotation info 14. Plenary Sessions of the Parliament of Finland a. Missing annotation info 15. CLAPI a. Missing size info b. Missing annotation info 16. Corpus de Français Parlé Parisien des années 2000 a. Missing size info b. Missing annotation info 17. Phonologie du Français Contemporain a. Missing size info b. Missing annotation info c. Missing licence info 18. GeWiss a. Missing size info b. Missing annotation info 19. Hamburg Adult Bilingual LAnguage (HABLA) a. Missing annotation info 20. CLIPS : corpora e lessici di italiano parlato e scritto a. Missing annotation info b. Missing licence info 21. Hamburg Corpus of Polish in Germany (HamCoPoliG) a. Missing annotation info 22. Consecutive and Simultaneous Interpreting (CoSi)	

a. Missing annotation info 23. ORAL2008: Balanced corpus of informal spoken Czech a. Missing annotation info 24. The Spoken Wikipedia Corpora a. Missing size info 25. Corpus of Spoken Estonian a. Missing licence info b. Missing annotation info c. Missing licence info 26. Belgische TV-Debatten a. Missing annotation info 27. Zweite Generation deutschsprachiger Migranten in Israel a. Missing annotation info 28. Hungarian Speecon Database a. Missing size info b. Missing annotation info 29. Mbochi speech corpus a. Missing annotation info 30. Corpus of Orleans a. Missing size info b. Missing annotation info	

2.7.3. Changelog

The following issues have been resolved and changes made for this resource family since the report on spoken corpora in the CLARIN infrastructure was published on 13 June 2018:

- 3 new corpora were added to the overview:
  1. The Icelandic Spoken Language Corpus
  2. Nordic Dialect Corpus v. 4.0
  3. Corpus of Orleans
- 1 corpus now has a VLO entry: LIA
- Size was resolved for the following corpus: Skolt Saami Documentation Corpus (2016)
- Licence has been added for 31 corpora:
  1. IFA Spoken Language Corpus (CLARIN PUB)
  2. LIA (CLARIN ACA)
  3. Deutsche Mundarten: Zwirner-Korpus (CLARIN RES)
  4. Deutsche Mundarten: ehemalige deutsche Ostgebiete (CLARIN RES)
  5. Deutsche Mundarten: DDR (CLARIN RES)
  6. Deutsche Umgangssprachen: Pfeffer-Korpus (CLARIN RES)
  7. Deutsche Standardsprache: König-Korpus (CLARIN RES)
  8. Deutsche Hochlautung (CLARIN RES)
  9. Australiendeutsch (CLARIN RES)
  10. Russlanddeutsche Dialekte (CLARIN RES)
  11. Emigrantendeutsch in Israel (CLARIN RES)
  12. Emigrantendeutsch in Israel: Wiener in Jerusalem (CLARIN RES)
  13. Forschungs- und Lehrkorpus gesprochenes Deutsch (CLARIN RES)
  14. Zweite Generation deutschsprachiger Migranten in Israel (CLARIN RES)
  15. Gesprochene Wissenschaftssprache Kontrastiv (CLARIN RES)
  16. Grundstrukturen: Freiburger Korpus (CLARIN RES)

    17. Dialogstrukturen (CLARIN RES)
    18. Berliner Wendekorpus (CLARIN RES)
    19. Biographische und Reiseerzählungen (CLARIN RES)
    20. Belgische TV-Debatten (CLARIN RES)
    21. Elizitierte Konfliktgespräche (CLARIN RES)
    22. Mehrsprachige Kinder im Vorschulalter (CLARIN RES)
    23. Gothenburg Dialogue Corpus (CC-BY)
    24. Corpus of Doctor-Patient Conversations from Ahus (CLARIN ACA)
    25. Corpus of Lecture Speech (CC-BY-SA)
    26. The BigBrother Corpus (CLARIN ACA)
    27. Corpus of American Nordic Speech (CANS) (CLARIN ACA)
    28. TAUS (CLARIN ACA)
    29. NoTa-Oslo (CLARIN ACA)
    30. Corpus of Radio Interviews (CC-BY)
    31. Corpus of Lecture Speech (CC-BY-SA)

- Annotation info has been added for the following 4 corpora:
    1. ALCEBLA (orthographic and phonetic transcription)
    2. Faroese Danish Corpus Hamburg 0.2.dan (FADAC-0.2.dan Hamburg) (EXMARaLDA annotation, described here)
    3. Nganasan Spoken Language Corpus (NSLC) (alignment of transcriptions and audio recordings)
    4. Prague DaTabase of Spoken Czech 1.0 (MSD-tagged, lemmatised)
- IFA speech corpus seems to be a duplicate variant of IFA Spoken Language Corpus, so it was removed from the resource family.

2.8. Manually annotated corpora

2.8.1. Summary

In Table 15, we summarize the information on manually annotated corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, annotation and licence. The summary is based on the manually annotated corpora subpage of the Resource Families that was last updated on 30 October 2020.

Table 15: Summary of information on manually annotated corpora within the CLARIN infrastructure

Identification	• 73 manually annotated corpora part of the CLARIN infrastructure in total • 69 (93%) corpora identified through the VLO • 4 (7%) corpora identified through national repositories, but not through VLO
Availability	• 19 (26%) corpora for download and through a concordancer • 51 (70%) corpora for download • 3 (4%) corpora through a concordancer
Languages	• 62 (85%) corpora are monolingual, 11 (15%) are multilingual • 8 Czech, 6 Estonian, 7 Polish, 7 Slovenian, 5 Portuguese, 3 Croatian, 3 English, 3 Finnish, 3 German, 2 Greek, 2 Dutch, 2 Hungarian, 2 Lithuanian, 2 Serbian corpora • 1 corpus per language: Arabic, Bulgarian, Danish, French, Icelandic, Norwegian, Tamil
Size	• Information on size missing for 1 (1%) corpus • Largest corpus has 407.5 million tokens, smallest has 22,000 tokens • 41 small corpora (<1 million words/tokens) • 14 medium-sized corpora (1–10 million words/tokens) • 3 large corpora (>10 million tokens)
Annotation	• 14 (19%) for PoS/MSD-tagging but not syntactic parsing • 7 (10%) for lemmatization • 35 (48%) for syntactic parsing • 11 (15%) for named entity recognition • 7 (10%) (for sentiment analysis • 26 (36%) for other annotation layers
Licence	• Information on licence available for all corpora • 46 (63%) corpora under CC-BY

2.8.2. Issues

In the following Table, we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 16: Issues

List of manually annotated corpora with metadata issues	List of CLARIN manually annotated corpora not in the VLO
1. Grundtvig's Works Corpus a. Missing size info	1. NKJP1M 2. Polish Coreference Corpus 3. Polish Dependency Bank in Universal Dependency format 4. Polish Summaries Corpus

2.8.3. Changelog

The following change has been made for this resource family since the report on manually annotated corpora in the CLARIN infrastructure was published on 11 March 2019:

- 1 new corpus has been added: Icelandic Parsed Historical Corpus (IcePaHC)
- 1 corpus is now findable through the VLO: Icelandic Parsed Historical Corpus (IcePaHC)
- Licence has been resolved for 2 corpora:
  1. Szeged Corpus 2.0 (Licence agreement)
  2. Szeged Treebank 2.0 (Licence agreement)

2.9. Literary corpora

2.9.1. Summary

In Table 17, we summarize the information on literary corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, annotation and licence. The summary is based on the literary corpora subpage of the Resource Families that was current on 30 October 2020.

Table 17: Summary of information on literary corpora within the CLARIN infrastructure

Identification	• 43 literary corpora part of the CLARIN infrastructure in total • 39 (91%) corpora identified through the VLO • 4 (9%) corpora identified through national repositories, but not through VLO
Availability	• 6 (14%) corpora for download and through a concordancer • 12 (28%) corpora for download • 22 (51%) corpora through a concordancer • 3 (7%) corpora unavailable
Languages	• 35 (81%) corpora are monolingual, 8 (19%) are multilingual • 8 Finnish corpora, 4 Polish corpora, 4 Norwegian corpora, 3 Swedish corpora, 3 Estonian corpora, 2 English corpora, 2 Greek corpora, 2 Spanish corpora • 1 corpus per language: Croatian, Danish, French, Latvian, Sami, Portuguese, Sumerian
Size	• Information on size missing for 7 (16%) corpora • Largest corpus has 34.5 million tokens, smallest has 4000 tokens • 8 small corpora (<1 million words/tokens) • 17 medium-sized corpora (1–10 million words/tokens) • 2 large corpora (>10 million words/tokens)
Annotation	• 10 (23%) corpora syntactically parsed, 7 (16%) corpora PoS/MSD-tagged • Unknown for 24 (56%) corpora
Licence	• 17 (40%) corpora under CC-BY • Unknown for 8 (19%) corpora

2.9.2. Issues

In the following Table, we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 18: Issues

List of literary corpora with metadata issues	List of CLARIN literary corpora not in the VLO
1. One-million Corpus of Croatian Literary Language a. Unavailable due to broken link b. Missing annotation info c. Missing licence info 2. Johannes V. Jensen Corpus a. Missing annotation info 3. Complete Corpus of Anglo-Saxon Poetry a. Missing size info b. Missing licence info 4. Collection of older original Estonian-language works of fiction a. Missing annotation info 5. Corpus of Estonian fiction a. Missing annotation info	1. aformes 2. République-Bastille (1948-1949) 3. Greek Medieval Texts 4. Cultural Thesaurus of the Greek Language

6. Estonian Runic Songs' Database
    a. Missing annotation info
7. Classics of Finnish Literature, Kielipankki Version
    a. Missing annotation info
8. Corpus of Old Literary Finnish
    a. Missing annotation info
    b. Missing licence info
9. Corpus of Early Literary Finnish
    a. Missing size info
    b. Missing annotation info
    c. Missing licence info
    d. Unavailable due to incorrect link
10. Corpus of Finnish Literary Classics
    a. Missing annotation info
    b. Missing licence info
11. The Finnish Gutenberg Corpus
    a. Missing annotation info
12. République-Bastille (1948-1949)
    a. Missing annotation info
13. Greek Medieval Texts
    a. Missing annotation info
14. Latvian literature classics
    a. Missing size info
    b. Missing annotation info
    c. Missing licence info
    d. Unavailable due to broken link
15. North Saami Corpus (Literature) (UHLCS)
    a. Missing annotation info
16. 1000 Novels Corpus
    a. Missing annotation info
17. 1000PLUS Novels Corpus (1.0)
    a. Missing annotation info
18. Late 19th- and Early 20th-Century Polish Novels
    a. Missing size info
    b. Missing annotation info
19. POE: Microcorpus of 20th century Polish poetry
    a. Missing size info
    b. Missing annotation info
20. LT Corpus
    a. Missing annotation info
21. Electronic corpus of 15th-century Castilian cancionero manuscripts
    a. Missing size info
    b. Missing annotation info
    c. Mssing licence info
22. Electronic text corpus of Sumerian literature (ETCSL)
    a. Missing annotation info

b. Missing licence info 23. Finnish Folk Poetry a. Missing annotation info 24. Classics Library of the National Library of Finland - Kielipankki version a. Missing size info b. Missing annotation info 25. aformes a. Missing annotation info	

2.9.3. Changelog

Since the report was published on 26 September 2019, 1 new corpus has been added:

(1) Cultural Thesaurus of the Greek Language

2.10. Academic corpora

2.10.1. Summary

In Table 19, we summarize the information on academic corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, annotation and licence. The summary is based on the academic corpora subpage of the Resource Families that was current on 30 October 2020.

Table 19: Summary of information on academic corpora within the CLARIN infrastructure

Identification	• 22 academic corpora part of the CLARIN infrastructure in total • 19 (86%) corpora identified through the VLO • 3 (14%) corpora identified through national repositories, but not through VLO
Availability	• 5 (23%) corpora for download and through a concordancer • 7 (32%) corpora for download • 9 (41%) corpora through a concordancer • 1 (4%) corpora unavailable
Languages	• 20 (91%) corpora are monolingual, 2 (9%) are multilingual • 5 English corpora, 3 Greek corpora, 3 Swedish corpora, 2 French corpora • 1 corpus per language: Czech, Estonian, Finnish, German, Russian, Slovenian, Spanish
Size	• Information on size available for all corpora • Largest corpus has 1.7 billion tokens, smallest has 48,000 tokens • 14 small corpora (<10 million words/tokens) • 5 medium-sized corpora (>10-100 million words/tokens) • 3 large corpora (>100 million words/tokens)
Annotation	• 7 (32%) corpora PoS/MSD-tagged, 5 (23%) corpora lemmatized • Unknown for 12 (55%) corpora
Licence	• 16 (73%) corpora under CC-BY • Information on licence available for all corpora

2.10.2. Issues

In the following Table, we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 20: Issues

List of academic corpora with metadata issues	List of CLARIN academic corpora not in the VLO
1. Czech Sociological Review a. Missing annotation info	1. OROSSIMO Corpus

2. UH's English E-thesis corpus a. Missing annotation info 3. Corpus of Estonian scientific texts a. Missing annotation info b. Unavailable due to unknown reasons 4. Chambers-Le Baron Corpus of Research Articles a. Missing annotation info 5. UH's French E-thesis corpus a. Missing annotation info 6. UH's German E-thesis corpus a. Missing annotation info 7. Modern Greek Dialects: scientific papers a. Missing annotation info 8. The Language of Literature and the Language of Translation (collected scientific papers) a. Missing annotation info 9. UH's Russian E-thesis corpus a. Missing annotation info 10. UH's Spanish E-thesis corpus a. Missing annotation info 11. Academic texts - humanities a. Missing annotation info 12. UH's Swedish E-thesis corpus a. Missing annotation info	2. Modern Greek Dialects: scientific papers 3. The Language of Literature and the Language of Translation (collected scientific papers)

2.10.3. Changelog

The following change has been made for this resource family since the report on academic corpora in the CLARIN infrastructure was published on 12 February 2020:

- 1 new corpus has been added: The Royal Society Corpus

2.11. Reference corpora

2.11.1. Summary

In Table 21, we summarize the information on reference corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, annotation and licence. The summary is based on the reference corpora subpage of the Resource Families that was current on 26 October 2020.

Table 21: Summary of information on reference corpora within the CLARIN infrastructure

Identification	• 30 reference corpora part of the CLARIN infrastructure in total • 20 (67%) corpora identified through the VLO • 10 (23%) corpora identified through national repositories, but not through VLO
Availability	• 10 (23%) corpora for download and through a concordancer • 5 (17%) corpora for download • 14 (47%) corpora through a concordancer • 1 (3%) corpora unavailable
Languages	• 29 (97%) corpora are monolingual, 1 (3%) is multilingual • 4 Slovenian, 3 Czech, 3 Greek, 2 Croatian, 2 English, 2 Estonian, 2 Norwegian corpora • 1 corpus per language: Abkhaz, Bulgarian, Danish, Dutch, German, Hungarian, Icelandic, Lithuanian, Polish, Portuguese, Welsh
Size	• Information on size available for all corpora • Largest corpus has 31.7 billion tokens, smallest has 3 million tokens • 1 very small corpora (<10 million words/tokens) • 9 small corpora (10–100> million words/tokens) • 14 medium-sized corpora (100–1,000 million words/tokens) • 6 large corpora (>1,000 million words/tokens)
Annotation	• 24 (80%) corpora PoS/MSD-tagged, 22 (73%) corpora lemmatized • Unknown for 4 (13%) corpora
Licence	• 8 (27%) corpora under CC-BY, 5 under CLARIN RES • Information on licence missing for 3 (10%) corpora

2.11.2. Issues

In the following Table, we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 22: Issues

List of reference corpora with metadata issues	List of CLARIN reference corpora not in the VLO
1. Croatian National Corpus a. Missing annotation info b. Missing licence info 2. Hungarian National Corpus a. Missing licence info 3. National Corpus of Polish a. Missing licence info 4. Hungarian National Corpus a. Missing annotation info 5. National Corpus of Polish a. Missing annotation info 6. CorCenCC a. Missing annotation info	1. AbNC: Abkhaz National Corpus 2. GNC: Georgian National Corpus 3. Norsk Ordboks Nynorskkorpus (NNK) 4. Hellenic National Corpus 5. Corpus of Greek Texts 6. Diachronic corpus of Greek of the 20th century 7. Written corpus Gigafida 2.0 8. Written corpus Kres 1.0 9. Bulgarian National Reference Corpus (BNRC) 10. CorCenCC

2.11.3. Changelog

1 new corpus has been added since the report on reference corpora was published on 23 July 2020: CorCenCC.

2.12. Multimodal corpora

In Table 23, we summarize the information on multimodal corpora in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, through a concordancer, or both), size, annotation and licence. The summary is based on the multimodal corpora subpage of the Resource Families that was current on 30 October 2020.

2.12.1. Summary

Table 23: Summary of information on multimodal corpora within the CLARIN infrastructure

Identification	• 16 multimodal corpora part of the CLARIN infrastructure in total • 15 (94%) corpora identified through the VLO • 1 (6%) corpora identified through national repositories, but not through VLO
Availability	• 1 (6%) corpora for download and through a concordancer • 12 (75%) corpora for download • 1 (6%) corpora through a concordancer • 2 (13%) corpora unavailable
Languages	• 12 (75%) corpora are monolingual, 4 (25%) is multilingual • 3 German corpora • 1 corpus per language: Dutch, English, Estonian, Finnish, French, Greek, Hungarian, Slovenian, and Zulu
Size	• Information on size is missing for 6 (37%) of all corpora • Largest corpus has 1.7 billion tokens, smallest has 48,000 tokens • 4 small corpora (<10 hours of recordings) • 3 medium-sized corpora (10–50 hours of recordings) • 2 large corpora (>50 hours of recordings)
Annotation	• 4 (25%) corpora annotated for gestures, 3 (19%) for gaze direction • Unknown for 5 (31%) corpora
Licence	• 5 (31%) corpora under CC-BY • Information on licence is missing for 4 (25%) corpora

2.12.2. Issues

In the following Table, we list the corpora that have missing metadata and those that cannot be found through the VLO.

Table 24: Summary of information on multimodal corpora within the CLARIN infrastructure

List of multimodal corpora with metadata issues	List of CLARIN multimodal corpora not in the VLO
1. Eye-tracking in Multimodal Interaction Corpus a. Missing size info b. Missing annotation info 2. SmartWeb Video Corpus (SVC) a. Missing size info 3. Unisa isiZulu Video Corpus a. Missing size info 4. Video-linked Thai/Swedish child data corpus a. Missing size info b. Missing licence info 5. MPI ESF Corpus a. Missing size info	1. Multimodal and multiparty corpus of text comprehension interactions

b. Missing annotation info c. Missing licence info 6. Multimodal and multiparty corpus of text comprehension interactions a. Missing size info 7. TV News Corpus a. Missing annotation info 8. Hindi Visual Genome 1.0 a. Missing annotation info 9. Unisa isiZulu Video Corpus a. Missing annotation info b. Missing licence info 10. Corpus d'interactions dialogales a. Missing licence info	

2.12.3. Changelog

No issues have been resolved or changes made for this resource family since the report on multimodal corpora in the CLARIN infrastructure was published on 9 September 2020.

3. Lexical resources

3.1. Lexica

3.1.1. Summary

In Table 25, we summarize the information on lexica in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, online browsing, or both), size, and licence. The summary is based on the lexica subpage of the Resource Families that was current on 4 November 2020.

Table 25: Summary of information on lexica within the CLARIN infrastructure

Identification	• 75 lexica part of the CLARIN infrastructure in total • 72 (96%) lexica identified through the VLO • 3 (4%) lexica identified through national repositories, but not through VLO
Availability	• 16 (21%) lexica for download and online browsing • 52 (69%) lexica for download • 7 (9%) lexica unavailable
Languages	• 62 (83%) lexica are monolingual, 13 (17%) are multilingual • 11 Portuguese, 8 Slovenian, 7 Dutch, 5 Czech, 5 French, 4 Italian, 4 Polish, 4 Swedish, 2 Arabic, 2 Croatian, 2 Danish, 2 English, 2 Serbian lexica • 1 lexicon per language: Estonian, Greek, Icelandic, Maltese
Size	• Information on size missing for 2 (3%) lexica • 23 small lexica (<10,000 entries) • 29 medium-sized lexica (10,000–100,000 entries) • 21 large lexica (>100,000 lexica)
Licence	• 45 (60%) lexica under CC-BY • Unknown for 1 (13%) lexicon

3.1.2. Issues

In the following Table, we list the lexica that have missing metadata and those that cannot be found through the VLO.

Table 26: Issues

List of lexica with metadata issues	List of CLARIN lexica not in the VLO
1. Basilex Lexicon a. Missing size info (strictly speaking, unclear) b. Missing licence info 2. Basiscript Lexicon a. Missing size info 3. VfrLPL a. Missing licence info	1. A machine-readable Persian-English dictionary 2. A machine-readable glossary of Egyptian Arabic 3. A machine-readable Persian-English glossary of verbs

3.1.3. Changelog

The following issues have been resolved and changes made for this resource family since the report on lexical resource in the CLARIN infrastructure was published on 23 July 2019:

- 1 lexicon now has a VLO entry: A machine-readable dictionary of Egyptian Arabic.
- Curiously, the other 3 Persian/English/Arabic lexica listed in Table 22, which are part of the same resource collection (Dictionary Gate), do not have VLO entries.

3.2. Dictionaries

3.2.1. Summary

In Table 27, we summarize the information on dictionaries in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, online browsing, or both), size, and licence. The summary is based on the dictionaries subpage of the Resource Families that was last updated on 30 October 2020.

Table 27: Summary of information on dictionaries within the CLARIN infrastructure

Identification	• 95 dictionaries part of the CLARIN infrastructure in total • 83 (87%) dictionaries identified through the VLO • 12 (13%) dictionaries identified through national repositories, but not through VLO
Availability	• 44 (47%) dictionaries for download and browsing • 18 (18%) dictionaries for download • 32 (34%) dictionaries for browsing • 1 (1%) dictionary unavailable
Languages	• 57 (60%) dictionaries are monolingual, 38 (40%) are multilingual • 20 Estonian, 9 Slovenian, 5 Dutch, 5 Latvian, 3 Arabic, 3 Finnish, 2 Polish, 2 Swedish • 1 dictionary per language: Dagaare, Danish, English, Frisian, German, Icelandic, Karelian, Veps
Size	• Information on size missing for 3 (3%) dictionaries • 29 small dictionaries (<10,000 entries) • 41 medium-sized dictionaries (10,000–100,000 entries) • 15 large dictionaries (>100,000 entries)
Licence	• 40 (42%) dictionaries under CC-BY • Unknown for 19 (20%) dictionaries

3.2.2. Issues

In the following Table, we list the dictionaries that have missing metadata and those that cannot be found through the VLO.

Table 28: Issues

List of dictionaries with metadata issues	List of CLARIN dictionaries not in the VLO
Information on size is missing for the following 3 dictionaries: 1. Multilingual dictionaries 2. Online Lexicon of Veps Language 3. The Dictionary of Estonian Dialects (2013-...) Information on licence is missing for the following 18 dictionaries: 1. Dictionary of Contemporary Dutch (ANW) 2. Dictionary of Middle Dutch (MNW) 3. Dictionary of Modern Latvian (MLVV) 4. Dictionary of Standard Latvian Language (LLVV) 5. Digital Dictionary of the German Language (DWDS) 6. English-Latvian Dictionary 7. Estonian-Latvian Dictionary	1. A machine-readable dictionary of Dagaare 2. Dictionary of Modern Latvian (MLVV) 3. English-Latvian Dictionary 4. Historical Dictionary of Latvian (16th-17th Centuries) (LVVV) 5. Latvian-Latgalian Dictionary 6. Latvian-Russian Dictionary 7. Lithuanian-Latvian Dictionary 8. Lithuanian-Latvian-Latgalian Dictionary (LLL) 9. Mühlenbach-Endzelin Latvian Dictionary (MEV) 10. Rendering Dictionary of Personal Names 11. Russian-Latvian Dictionary

8. Historical Dictionary of Latvian (16th-17th Centuries) (LVVV) 9. Latvian-English Dictionary 10. Latvian-Latgalian Dictionary 11. Latvian-Lithuanian Dictionary 12. Latvian-Russian Dictionary 13. Lithuanian-Latvian Dictionary 14. Lithuanian-Latvian-Latgalian Dictionary (LLL) 15. Livonian-Estonian-Latvian Dictionary 16. Mühlenbach Endzelin Latvian Dictionary (MEV) 17. Rendering Dictionary of Personal Names 18. Russian-Latvian Dictionary	

3.2.3. Changelog

The following issues have been resolved and changes made for this resource family since the report on lexical resource in the CLARIN infrastructure was published on 23 July 2018:

- 1 new dictionary has been added: Anglų kalbos žodynas
- The following 3 dictionaries now have VLO entries:
    1. A machine-readable dictionary of Damascus Arabic
    2. A machine-readable dictionary of Modern Standard Arabic
    3. A machine-readable dictionary of Tunis Arabic
- Licence info has been added for the following dictionary:
    1. Tezaurs.lv (CC-BY)

3.3. Conceptual resources

3.3.1. Summary

In Table 29, we summarize the information on conceptual resources in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, online browsing, or both), size, and licence. The summary is based on the conceptual resources subpage of the Resource Families that was current on 30 October 2020.

Table 29: Summary of information on conceptual resources within the CLARIN infrastructure

Identification	• 29 conceptual resources part of the CLARIN infrastructure in total • 29 (100%) conceptual resources identified through the VLO
Availability	• 7 (24%) conceptual resources for download and browsing • 17 (59%) conceptual resources for download • 3 (10%) conceptual resources for browsing • 2 (7%) conceptual resources unavailable
Languages	• 22 (76%) conceptual resources are monolingual, 7 (24%) are multilingual • 4 Swedish, 3 Italian, 3 Portuguese, 2 Danish, 2 Polish conceptual resources • 1 conceptual resource per language: Ancient Greek, Dutch, Estonian, Finnish, German, Greek, Maltese, Slovenian
Size	• Information on size is available for all conceptual resources • 9 small conceptual resources (<10,000 entries) • 11 medium-sized conceptual resources (10,000–100,000 entries) • 9 large conceptual resources (>100,000 entries)
Licence	• 21 (72%) conceptual resources under CC-BY • Unknown for 1 (3%) conceptual resources

3.3.2. Issues

In the following Table, we list the conceptual resources that have missing metadata and those that cannot be found through the VLO.

Table 30: Issues

List of conceptual resources with metadata issues	List of CLARIN conceptual resources not in the VLO
1. IWN-LOD a. Missing licence info	N/A

3.3.3. Changelog

No issues have been resolved or changes made for this resource family since the report on lexical resource in the CLARIN infrastructure was published on 23 July 2018.

3.4. Glossaries

3.4.1. Summary

In Table 31, we summarize the information on glossaries in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, online browsing, or both), size, and licence. The summary is based on the glossaries subpage of the Resource Families that was last updated on 30 October 2020.

Table 31: Summary of information on glossaries within the CLARIN infrastructure

Identification	• 32 glossaries part of the CLARIN infrastructure in total • 32 (100%) glossaries identified through the VLO
Availability	• 7 (22%) glossaries for download and browsing • 20 (62%) glossaries for download • 5 (16%) glossaries for browsing
Languages	• 10 (31%) glossaries are monolingual, 22 (69%) are multilingual • 3 Estonian, 2 English, 2 Greek glossaries • 1 glossary per language: Dutch, Slovenian, Swedish
Size	• Information on size missing for 3 (9%) glossaries • 16 small glossaries (<10,000 entries) • 5 medium-sized glossaries (10,000–100,000 entries) • 5 large glossaries (>100,000 entries)
Licence	• 23 (72%) glossaries under CC-BY • Unknown for 1 (3%) glossary

3.4.2. Issues

In the following Table, we list the glossaries that have missing metadata and those that cannot be found through the VLO.

Table 32: Issues

List of spoken glossaries with metadata issues	List of CLARIN spoken glossaries not in the VLO
The following 3 glossaries lack information on size: 1. Multilingual glossary (Department of Foreign Languages, Translation and Interpretation) 2. Name Component Lexicon 3. The Family Name Database of the Institute of the Estonian Language The following glossary does not specify the licence (though it does specify conditions of use): 1. Λεξικό Γλωσσολογικών όρων: Γερμανικά – Ελληνικά - Αγγλικά (lexicon of linguistic terms: DE-EL-EN)	N/A

3.4.3. Changelog

No issues have been resolved or changes made for this resource family since the report on lexical resource in the CLARIN infrastructure was published on 23 July 2018.

3.5. Wordlists

3.5.1. Summary

In Table 33, we summarize the information on wordlists in terms of their identification (i.e. found through VLO or only through a national repository), availability (for download, online browsing, or both), size, and licence. The summary is based on the wordlists subpage of the Resource Families that was current on 30 October 2020.

Table 33: Summary of information on wordlists within the CLARIN infrastructure

Identification	• 53 wordlists part of the CLARIN infrastructure in total • 52 (98%) wordlists identified through the VLO • 1 (2%) wordlists identified through national repositories, but not through VLO
Availability	• 2 (4%) wordlists for download and for browsing • 45 (85%) wordlists for download • 6 (11%) wordlists for browsing
Languages	• 29 (55%) wordlists are monolingual, 24 (45%) are multilingual • 9 Finnish, 4 Slovenian, 4 Swedish, 3 Estonian, 2 Dutch, 2 Maltese wordlists • 1 wordlist per language: German, Greek, Ngbugu, Polish, Spanish
Size	• Information on size missing for 6 (11%) wordlists • 24 small wordlists (<10,000 entries) • 10 medium-sized wordlists (10,000–100,000 entries) • 11 large wordlists (>100,000 entries)
Licence	• 38 (72%) wordlists under CC-BY • Unknown for 4 (8%) wordlists

3.5.2. Issues

In the following Table, we list the wordlists that have missing metadata and those that cannot be found through the VLO.

Table 34: Issues

List of wordlists with metadata issues	List of CLARIN wordlists not in the VLO
The following wordlists lack information on size: 1. Dictionary for the CST Lemmatizer 2. JRC-Names - a multilingual named entity resource 3. Labial vibrants in Mangbetu: Archival form 4. Names of Countries 5. The Conceptual File of Estonian Lexis of the Institute of Estonian Language 6. Vocabulaire de Pathologies humaines The following 4 wordlists lack information on licence: 1. Lexical functions of Spanish verb-noun collocations 2. Neologisms Online v3 3. Ngbugu digital wordlist: Archival form 4. Deutsche Wortschatz	1. Deutsche Wortschatz

3.5.3. Changelog

The following issue has been resolved for this resource family since the report on lexical resource in the CLARIN infrastructure was published on 23 July 2018:

- Size info has been added for the following wordlist: Vocabulaire de Transferts de chaleur (1462 entries)

4. Tools

4.1. Tools for normalization

4.1.1. Summary

In Table 35, we summarize the information on text normalizers in terms of their identification (i.e. found through VLO or only through a national repository/webpage), availability (for download, as a web application, or both), size, functionality, input and output, and licence. The summary is based on the tools for normalization subpage of the Resource Families that was last updated on 4 November 2020.

Table 35: Summary of information on text normalizers within the CLARIN infrastructure

Identification	• 14 normalizers part of the CLARIN infrastructure in total • 4 (29%) normalizers identified through the VLO • 10 (71%) normalizers identified through national repositories, but not through VLO
Availability	• 2 (14%) normalizers downloadable and used as a web application • 5 (36%) normalizers downloadable • 3 (21%) normalizers used as a web application • 4 (29%) normalizers unavailable
Languages	• 11 (79%) normalizers for a single language, 3 (21%) multilingual • 3 normalizers for Dutch, 3 for German • 1 normalizer per language: English, Hungarian, Icelandic, Slovenian, Turkish
Functionality	• 7 (50%) tools perform normalization only • 7 (50%) tools perform additional tasks, such as PoS/MSD-tagging (6), lemmatization (5), and named entity recognition (4)
Input/output	• 5 (36%) tools do not specify input and output type or only partially
Licence	• 5 (36%) normalizers available under GNU GPL • Unknown for 7 (50%) normalizers

4.1.2. Issues

In the following Table, we list the tools that have missing metadata and those that cannot be found through the VLO.

Table 36: Issues

List of normalizers with metadata issues	List of normalizers not in the VLO
1. CSMTiser a. Missing licence info 2. Normo a. Unspecified input and output format b. Missing licence info 3. Skrambi a. Unspecified input and output format b. Missing licence info 4. CAB orthographic canonicalizer a. Unspecified output format info b. Missing licence info 5. CAB historical text analysis a. Unspecified input and output format	1. Text Tonsorium 2. FoLiA-wordtranslate 3. Nederlab Pipeline 4. TiCClops 5. @Philostei 6. PICCL: Philosophical Integrator of Computational and Corpus Libraries 7. VARD2 8. DTA::CAB 9. Normo 10. Skrambi 11. CSMTiser

b. Missing licence info 6. @Philostei a. Missing licence info 7. TiCClops a. Missing licence info 8. Turkish Natural Language Processing Pipeline a. Missing licence info	

4.1.3. Changelog

No issues have been solved since the report on text normalizers was published on 12 December 2019.

4.2. Tools for named entity recognition

4.2.1. Summary

In Table 37, we summarize the information on named entity recognizers in terms of their identification (i.e. found through VLO or only through a national repository/webpage), availability (for download, as a web application, or both), size, functionality, NE categories recognized, and licence. The summary is based on the tools for named entity recognition subpage of the Resource Families that was last updated on 30 October 2020.

Table 37: Summary of information on named entity recognizers within the CLARIN infrastructure

Identification	• 24 NERs part of the CLARIN infrastructure in total • 19 (79%) NERs identified through the VLO • 5 (21%) NERs identified through national repositories, but not through VLO
Availability	• 5 (21%) NERs downloadable and used as a web application • 8 (33%) NERs downloadable • 10 (42%) NERS used as a web application • 1 (4%) NER unavailable
Languages	• 15 (63%) NERs for a single language, 9 (37%) multilingual • 4 Dutch, 3 Polish, 2 English, 2 German NERs • 1 NER per language: Finnish, Greek, Hungarian, Portuguese
Functionality	• 16 (67%) tools are dedicated NERs • 8 (33%) tools also perform PoS/MSD-tagging
NER categories	• Unspecified for 6 (25%) NERs • 18 NERs recognize person, 17 organisation and location, 6 date and time
Licence	• 6 (25%) NERs under GPL • Unknown for 5 (21%) NERs

4.2.2. Issues

In the following Table, we list the tools that have missing metadata and those that cannot be found through the VLO.

Table 38: Issues

List of NERs with metadata issues	List of NERs not in the VLO
1. The NERD named entity recognizer a. Does not specify NE categories b. Missing licence info 2. hunner - named entitiy recognizer for Hungarian a. Does not specify NE categories b. Missing licence info 3. Nerf a. Does not specify NE categories 4. NameScape: Named Entity Recognition a. Does not specify NE categories b. Missing licence info 5. NCHLT Tagger a. Does not specify NE categories 6. CTexTools 2 a. Does not specify NE categories 7. LX-NER a. Does not specify licence info 8. INL labs a. Does not specify licence info	1. OpenNLP Name Finder (English) 2. GrNE-Tagger 3. Nerf 4. PolDeepNer 5. janes-ner

4.2.3. Changelog

No issues have been solved since the report on named entity recognizers was published on 23 January 2020.

4.3. Part-of-speech taggers and lemmatizers

4.3.1. Summary

In Table 39, we summarize the information on part-of-speech taggers and lemmatizers in terms of their identification (i.e. found through VLO or only through a national repository/webpage), availability (for download, as a web application, or both), size, functionality, input and output, and licence. The summary is based on the part-of-speech taggers and lemmatizers subpage of the Resource Families that was last updated on 4 November 2020.

Table 39: Summary of information on part-of-speech taggers and lemmatizers in the CLARIN infrastructure

Identification	• 66 tools part of the CLARIN infrastructure in total • 54 (82%) tools identified through the VLO • 12 (18%) tools identified through national repositories, but not through VLO
Availability	• 11 (17%) tools downloadable and used as a web application • 29 (44%) tools downloadable • 18 (28%) tools used as a web application • 7 (11%) tools unavailable
Languages	• 51 (77%) tools for a single language, 15 (23%) multilingual • 10 Bantu, 7 Polish, 5 German, 4 English, 4 Portuguese, 3 Dutch, 2 Afrikaans, 2 Estonian, 2 Slovenian tools • 1 tool per language: Assamese, Belarussian, Bulgarian, Czech, Finnish, Greek, Hungarian, Icelandic, Italian, Maltese, Norwegian
Functionality	• 23 (35%) tools perform part-of-speech tagging/morphosyntactic annotation only • 17 (26%) tools perform lemmatization only • 26 (39%) tools perform additional tasks such as syntactic parsing or named entity recognition
Input/output	• 41 (62%) tools do not specify input and output type or only partially
Licence	• 16 (24%) tools under GPL, 10 (15%) under GNU GPL • Unknown for 22 (33%) tools

4.3.2. Issues

In the following Table, we list the tools that have missing metadata and those that cannot be found through the VLO.

Table 40: Issues

List of taggers and lemmatizers with metadata issues	List of taggers and lemmatizers not in the VLO
1. NCHLT Sepedi Lemmatiser a. Missing input and output info 2. NCHLT Sesotho Lemmatiser a. Missing input and output info 3. NCHLT Setswana Lemmatiser a. Missing input and output info 4. NCHLT Siswati Lemmatiser a. Missing input and output info 5. NCHLT isiZulu Lemmatiser a. Missing input and output info 6. NCHLT isiXhosa Lemmatiser a. Missing input and output info 7. NCHLT isiNdebele Lemmatiser	1. ReLDIanno 2. janes-tagger 3. CLARIN DK NLP Toolbox 4. ILSP Feature-based multi-tiered POS Tagger 5. OpenNLP Part-of-Speech Tagger (English) 6. OpenNLP Part-of-Speech Tagger (German) 7. OpenNLP Part-of-Speech Tagger (Portuguese) 8. Sparv

   a. Missing input and output info

8. NCHLT Afrikaans Lemmatiser
   a. Missing input and output info
9. NCHLT Tshivenda Lemmatiser
   a. Missing input and output info
10. NCHLT Xitsonga Lemmatiser
    a. Missing input and output info
11. Tadpole
    a. Missing input and output info
    b. Missing licence info
12. Freeling
    a. Missing input and output info
    b. Missing licence info
13. LX-Tagger
    a. Missing input and output info
14. HMM tagger
    a. Missing input and output info
15. hunpos
    a. Missing input and output info
16. The Oslo-Bergen tagger
    a. Missing input and output info
17. Assamese POS Tagger
    a. Missing input and output info
    b. Missing licence info
18. Tagger SentiOne - version 2
    a. Missing input and output info
19. MorphoDiTa-based tagger for Polish language
    a. Missing input and output info
20. WCRFT (Wrocław CRF Tagger)
    a. Missing input and output info
21. TaKIPI
    a. Missing input and output info
    b. Missing licence info
22. Sepedi Part of Speech Tagger
    a. Missing input and output info
    b. Missing licence info
23. MLSS Tagger Web Service
    a. Missing input and output info
24. NLP-PIPE
    a. Missing input and output info
25. LX-Verbal Lemmatizer
    a. Missing input and output info
    b. Missing licence info
26. CST’s lemmatizer
    a. Missing input and output info
    b. Missing licence info
27. RFTagger
    a. Missing input and output info
28. GENIA Tagger
    a. Missing input and output info

9. CLaRK
10. NLP-PIPE
11. Turku-neural-parser-pipeline
12. CLAWS

29. STEPP Tagger
    a. Missing input and output info
30. NCHLT Tagger
    a. Missing input and output info
31. INL Labs tagger/lemmatizer tools
    a. Missing input and output info
32. SMOR lemmatizer
    a. Missing output format info
    b. Missing licence info
33. SepVerb Lemmatizer
    a. Missing output format info
    b. Missing licence info
34. MorphAdorner Lemmatizer
    a. Missing output format info
    b. Missing licence info
35. Character-level part-of-speech tagger of Slovene language
    a. Missing output format info
36. janes-tagger
    a. Missing output format info
37. Frog
    a. Missing input format info
38. WebLicht Part-of-Speech Tagger
    a. Missing output format info
    b. Missing licence info
39. Sticker part-of-speech tagger UD
    a. Missing input and output info
40. PoS Tagger OpenNLP Project
    a. Missing output format info
    b. Missing licence info
41. TreeTagger
    a. Missing input format info
42. janes-tagger
    a. Missing licence info
43. Tagger WS
    a. Missing licence info
44. Afrikaans TnT-Tagger
    a. Missing licence info
45. CLARIN DK NLP Toolbox
    a. Missing licence info
46. Sparv
    a. Missing licence info
47. CLaRK
    a. Missing licence info
48. Stanford Dependency Parser
    a. Missing licence info
49. Stanford Phrase Structure Parser
    a. Missing licence info
50. Stuttgart Dependency Parser
    a. Missing licence info
51. Lemmatizer

a. Missing licence info	

4.3.3. Changelog

Since the report on part-of-speech taggers and lemmatizers was published on 1 April 2020, the following 1 tool has been added: CLAWS.

4.4. Sentiment analyzers

In Table 41, we summarize the information on tools for sentiment analysis in terms of their identification (i.e. found through VLO or only through a national repository/webpage), availability (for download, as a web application, or both), size, functionality, and licence. The summary is based on the tools for sentiment analysis subpage of the Resource Families that was last updated on 30 October 2020.

4.4.1. Summary

Table 41: Summary of information on sentiment analyzers in the CLARIN infrastructure

Identification	• 5 tools part of the CLARIN infrastructure in total • 3 (60%) tools identified through the VLO • 2 (40%) tools identified through national repositories, but not through VLO
Availability	• 2 (40%) tools downloadable • 2 (40%) tools used as a web application • 1 (20%) tools unavailable
Languages	• 51 (77%) tools for a single language, 15 (23%) multilingual • 2 Polish tools • 1 tool per language: Greek and Finnish
Functionality	• 3 (60%) for opinion mining in addition to sentiment analysis • 2 (40%) for dependency subtree sentiment
Licence	• 2 (40%) under BSD 2 Clause licence, 1 (20%) under LGPL • Unknown for 1 (20%) tools

4.4.2. Issues

In the following Table, we list the tools that have missing metadata and those that cannot be found through the VLO.

Table 42: Issues

List of sentiment analyzers with metadata issues	List of sentiment analyzers not in the VLO
1. finsentiment (FIN-CLARIN) a. Missing licence info b. Missing input/output info 2. Etuma Customer Feedback Analysis a. Missing input/output info	1. finsentiment 2. Sentiment Analysis Tool

4.4.3. Changelog

No issues have been solved since the report on tools for sentiment analysis was published on 23 January 2020.

5. Summary

In this report, we have presented a detailed overview of all the current corpus, lexical, and tool families that are published under the CLARIN ERIC Resource Families initiative. For each family, we have summarized the salient metadata (inclusion in VLO, availability, linguistic annotation/output and input format, size, and licence) and drawn a comprehensive list of identified issues related to missing metadata. We have also described which changes have been made and which issues have been resolved since either the last report on corpus families issues, which was published in June 2018, or the individual family reports, provided the reports were published after the June 2018 report.

5.1. Corpora

Table 43 summarizes the salient features of the 12 corpus families, presenting information on the total number of corpora and the number of corpora with VLO entries, as well as the number of corpora with missing information on size, linguistic annotation, and licence.

Corpus family	Corpora	VLO		w/o size		w/o anno		w/o licence	
parliamentary	25	22	88%	0	0%	3	12%	3	12%
CMC	15	12	80%	0	0%	4	27%	2	13%
parallel	86	53	62%	11	13%	27	31%	6	7%
newspaper	33	22	67%	2	6%	14	42%	3	9%
L2-learner	74	72	98%	8	11%	20	27%	4	11%
historical	745	66	90%	12	16%	18	25%	4	5%
spoken	90	89	99%	12	13%	26	29%	5	6%
manually annotated	73	69	93%	1	1%	n/a	n/a	0	0%
literary	43	39	91%	7	16%	24	56%	8	19%
academic	22	19	86%	0	0%	12	55%	0	0%
reference	30	20	67%	0	0%	4	13%	3	10%
multimodal	16	15	94%	6	37%	5	31%	4	25%
Σ	581	498	86%	59	10%	157	27%	42	7%

Table 43: Summary of corpus families, their inclusion in the VLO, and number of metadata issues (size, annotation, and licence).

In total, there are 581 corpora across the 12 corpus families. The largest families are spoken corpora and parallel corpora, which consist of 89 and 86 corpora, respectively, while the smallest are CMC and multimodal corpora, which consist of 13 and 16 corpora, respectively. The vast majority of the corpora (498 or 86% out of 581) are findable in the VLO, where this share is lower for families with the smallest number of corpora with VLO entries (parallel and newspaper corpora) because they contain a number of resources that are either available through the CLARIN:EL repository or SWE-CLARIN's Språkbanken Text resource list, neither of which is as of yet harvested by the VLO. Information on size is overall readily available (missing only for 59 or 10% of all the corpora); moreover, it is included for all corpora within 4 families (parliamentary, CMC, academic, and reference corpora). Similarly, information on licence is missing 42 (7%) corpora overall; it is, however, available for all corpora within 1 family (multimodal corpora). Information on annotation fares the worst, as it is missing for 157 (27%) corpora overall and is in the case of 2 families – literary, and academic corpora – missing for more than half of the corpora in the family.

Table 44 summarizes several types of information regarding metadata issues:

i. the original number of metadata issues,
ii. the number of issues that have been solved since the previous 2018 corpus family report or the individual family reports published after said report,[3]
iii. the remaining number of issues,
iv. the number of corpora with remaining metadata issues,[4] and
v. the total number of corpora in the VLO and the number that originally did not have VLO entries but now have them.

Corpus family	Corpora	# original issues	# solved issues		# remaining issues	Current # corpora with issues		All VLO entries	New VLO entries	
parliamentary	25	10	4	40%	6	6	24%	22	1	5%
CMC	15	10	2	20%	8	5	33%	12	0	0%
parallel	86	52	8	15%	44	39	45%	53	1	2%
newspaper	33	22	3	14%	19	15	45%	22	0	0%
L2-learner	74	37	5	14%	32	23	31%	72	38	53%
historical	74	55	19	35%	36	29	40%	66	4	6%
spoken	90	79	36	47%	43	30	33%	89	1	1%
manually annotated	73	3	2	67%	1	1	1%	69	1	1%
literary	43	38	0	0%	38	25	58%	39	0	0%
academic	22	12	0	0%	12	12	55%	19	0	0%
reference	30	7	0	0%	7	6	20%	20	0	0%
multimodal	16	15	0	0%	15	10	63%	15	0	0%
Σ	581	340	79	23%	261	201	35%	498	46	9%

Table 44: Summary of metadata issues and their resolution for the corpus families

In sum, 340 issues pertaining to missing metadata were originally identified while a total of 79 (23%) of them have been solved. The most issues have been resolved for manually annotated corpora (2 or 67% of the originally identified issues), spoken corpora (36 or 47% of the originally identified issues) and parliamentary corpora (4 or 40% of the originally identified issues), while significantly fewer issues have been resolved for newspaper corpora (3 or 14%) and L2-learner corpora (5 or 14%). No issues have been solved for literary corpora, academic corpora, reference corpora, and multimodal corpora; however, what is important to note is that these are our most recent surveys which, except for literary corpora, have only been published in 2020. Finally, only 46 (9%) corpora represent new VLO entries since the last round of reporting – most new entries are within the L2-learner family, where all the 38 corpora with new VLO entries are TalkBank corpora.

[3] Cf. the individual changelogs for details.

[4] We provide this information in addition to the number of remaining issues because the latter number reflects the fact that a single corpus may display several issues; e.g., a corpus might lack information on both size and metadata.

5.2. Lexical resources

Table 45 summarizes the salient features of the 5 lexical resource families, presenting more or less the same information as above.

Lexical family	Resources	VLO		w/o size		w/o licence	
lexica	75	72	96%	2	3%	1	1%
dictionaries	95	83	87%	3	3%	19	20%
conceptual	29	29	100%	0	0%	1	3%
glossaries	32	32	100%	3	9%	1	3%
wordlists	53	52	98%	6	11%	4	8%
Σ	284	268	95%	14	5%	26	9%

Table 45: Summary of lexical resource families, their inclusion in the VLO, and number of metadata issues (size and licence)

In total, there are 284 resources across the 5 lexical families, where the largest family is dictionaries with 95 resources and the smallest conceptual resources with 29 resources. In comparison to the corpus families summarized in Table 43, lexical resources are slightly more readily found in the VLO, as 268 of them (95% vs. 86% in the case of the corpora) have VLO entries – the biggest outlier with the fewest number of VLO entries is the dictionaries family, where a set of 11 Latvian resources are not included in the VLO, presumably because they are not part of a certified repository whose data could be harvested by the VLO. The two types of metadata – size and licence – are all fairly readily included, with size not being included for a total of 14 (5%) resources and licence for 26 (9%) of the resources; in the case of licence, the outlier is again the dictionaries family, where most of the resources without licence are again Latvian dictionaries lacking repository entries, which in turn highlights the importance that resources be included in certified repositories conforming to e.g. FAIR principles.

Table 46 summarizes, in a similar way to Table 44 above, the changes that have been made for the 5 lexical resource families since the report on lexical resources was published in July 2019. While only 2 metadata issues have been resolved (5% of total issues; 1 missing licence in the case of dictionaries and 1 missing size in the case of wordlists), a total of 4 (1%) resources have now got VLO entries, which is a relatively high amount of issues solved given that, on the one hand, the lexical resource families represent more recent inclusions than most of the corpus families and, on the other, they report relatively small number of outstanding resources with issues to begin with.

Lexical family	Resources	# original issues	# solved issues		# remaining issues	Current # resources with issues		All VLO entries	New VLO entries	
lexica	75	3	0	0%	3	3	4%	72	1	1%
dictionaries	95	23	1	5%	22	22	23%	83	3	4%
conceptual	29	1	0	0%	1	1	3%	29	0	0%
glossaries	32	4	0	0%	4	4	13%	32	0	0%
wordlists	53	11	1	9%	10	10	19%	52	0	0%
Σ	284	42	2	5%	40	40	14%	268	4	1%

Table 46: Summary of metadata issues and their resolution for the lexical resource families

5.3. Tools

Finally, Table 47 summarizes the salient features of the 4 tool families, focusing on their inclusion in the VLO and the 2 types of metadata for which issues have been observed: input/output format and licence. In sum, there are 109 tools across the 4 families, with the largest family being part-of-speech taggers and lemmatizers (66 tools) and the smallest sentiment analyzers (5 tools). Unlike the corpus and lexical resource families summarized in Tables 43 and 45, the tools are very unevenly included in the VLO, with part-of-speech taggers and lemmatizers having VLO entries in 83% cases and normalizers having VLO entries only in 29% cases. Similarly, inclusion of the input and output formats as well as licence similarly varies between the tool families, where for instance 41 (63%) of part-of-speech taggers and lemmatizers do not include information on the input/output format, significantly fewer normalizers – 5 (36%) – lack this information. Since the tool families are by far the most recent inclusions in the entire Resource Families initiative, none of the issues listed in the report have been resolved thus far.

Tool family	Tools	VLO		w/o input/output		w/o licence	
norm	14	4	29%	5	36%	7	50%
NE	24	19	79%	n/a	n/a[5]	5	21%
PoS/lemma	66	54	83%	41	63%	22	34%
sentiment	5	3	60%	2	40%	1	20%
Σ	109	77	73%	48	44%	35	32%

Table 47: Summary of tool families, their inclusion in the VLO, and number of metadata issues (input/output format and licence); *norm* stands for (tools for) normalization, *NE* for named entities, *PoS* for part of speech tagging, *lemma* for lemmatization, and *sentiment* for sentiment analysis

[5] For Named Entity Recognizers, we have not yet surveyed the inclusion of the output and input formats but rather the inclusion of information on NE categories identified by the tools; cf. Section 4.2.2 for details.

6. Conclusion

In conclusion, there is now a total of 974 entries in CLARIN Resource Families. The surveyed corpora cover 30 different languages, with the most represented languages being English (55) and German (51). The overviews of lexical resources cover 24 different languages: with the most represented languages being Estonian (26), Slovenian (22) and Dutch (16). The featured tools cover 21 different languages, mostly Polish (12), Dutch (10) and German (10).

In addition to providing manually curated, user-friendly overviews of the resources and tools, the initiative has also resulted in a thorough evaluation of their records in our repositories, using the identified 384 issues pertaining to missing metadata on annotation, size, licence, or input/output formats as an opportunity to further improve the state of the infrastructures. 81 of the identified issues have already been solved since June 2018, when the first version of the report was published. Curation is in part facilitated through the GitHub page listing the issues which has also been updated the since the 2018 report. We have assigned to all the GitHub issues labels specifying the CLARIN consortium that is responsible or most closely associated with the issue, thereby further incentivizing individual CLARIN consortia to help with the curation process. Some issues have already been resolved through this channel and we would like to give our warm thanks to all the national CLARIN representatives who have already responded to the GitHub listings and helped curate the Families.

In 2021, we plan the following activities:

- In order to increase the usability of the resources and tools, we will form a special taskforce that will help us curate the existing resources and tools in the CLARIN infrastructure, and develop and implement preventive measures and prepare depositing guidelines which will minimize the number of metadata issues for any newly deposited resources and tools in the future. This is of crucial importance because existing issues hinder the reuse of any resources and tools in the CLARIN infrastructure but especially those featured in CLARIN Resource Families, which has proven to be a highly visible initiative appreciated by a broad spectrum CLARIN users and therefore warrants continued and careful upkeep.
- We will evaluate the resource and tool families from a more qualitative perspective, taking into account not only the availability or absence of metadata, but also in which way metadata are reported for the observed categories. For instance, it is often the case that resource size is reported in different ways for resources in the same resource family (e.g., corpus size in terms of sentences, tokens, words, hours or file size), which hinders cross-comparability of the resources. Certain resources also specify their annotation layers or licence information very imprecisely (e.g. using vague descriptors such as “multitagged” annotations and “other” licence), which is not very helpful for the users.
- We will promote a greater inclusion of key publications describing the tools and resources. Listing key publications is not only important from the perspective of ensuring and facilitating author attribution, but the publications themselves generally provide the most detailed descriptions of the resources and tools, thereby crucially complementing the metadata presented in CLARIN repositories with documentation that enable appropriate reuse of resources and tools as well as interpretation of research results.
- We will develop intensify dissemination activities the CRF initiative and proactively encourage depositing of existing impactful resources with CLARIN.
- We have published a designated call to fund small projects from the CLARIN network that can contribute to the CRF initiative with 3–6 PMs per project.
- We will perform a gap analysis for the CRF initiative and use the findings to steer further work within this initiative.