Title	Overview of reference corpora in the CLARIN infrastructure
Version	1.0
Author(s)	Darja Fišer, Jakob Lenardič
Date	23-07-2020
Status	For distribution
Distribution	BoD, NCF, UI
ID	CE-2020-1701

Table of contents

1. Introduction

In the following report, we present an overview of reference corpora that are part of the CLARIN infrastructure (i.e., they are either listed in the VLO or in the repositories of the national consortia). We understand reference corpora as corresponding to the following definition:

> A reference corpus is designed to provide comprehensive information about the language […] It has to be a general corpus of wide coverage of the language, and hopefully it will be treated by its user community as some kind of “standard” for the language. (Leech 2002)

Reference corpora thus contrast with specialised corpus families (e.g., parliamentary corpora, CMC-corpora) in that they are comprehensive with respect to genre inclusion, typically sampling a diverse set of primarily written genres.

The report was conducted in three steps:

(i) by manually searching the VLO and the national consortia with the following keywords: “national corpus”, “representative corpus”, “balanced corpus”, and “reference corpus”

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

The full results are available in this Google Docs Spreadsheet. In total, 38 corpora were identified in the initial survey. However, after reviewing the list, we found that only 26 of the identified corpora correspond to the above definition of reference corpora. The remaining 12 corpora mostly correspond to specialised corpora such as literary corpora or academic corpora. Furthermore, 9 corpora were already included in other CLARIN Resource Families, while the remaining 3 have now been added, so they were excluded from this report – see the second sheet in the Google Docs Spreadsheet for the reasons of exclusion for each of the 12 corpora. Finally, after an additional round of review of the previous version of this report by CLARIN UI and NC coordinators, 3 additional corpora have been added.

In Section 2, we first provide a comprehensive list of the reference corpora that are part of the CLARIN infrastructure. We subsequently describe their identification (i.e., listed in the VLO or not), their availability (download or through a concordancer), and their metadata (language, size, research discipline and period of publication, annotation, and license). Section 3 concludes the report with a discussion of the current situations and proposals for future improvements.

2. Reference corpora in the CLARIN infrastructure

29 reference corpora have been identified in the CLARIN infrastructure. In what follows, we group them according to language (monolingual and multilingual corpora) and describe them with the following metadata:

(i) Size
(ii) Annotation
(iii) Licence
(iv) Language
(v) Availability (online querying and/or download)
(vi) Publication

Table 1: Reference corpora in the CLARIN infrastructure, sorted by language

Corpus	Language	Description
AbNC: Abkhaz National Corpus **Size:** 10 million words **Annotation:** MSD-tagged, lemmatized **Licence:** CLARIN_PUB-BY-NC-ND	Abkhaz	This corpus includes Abkhaz texts published between 1920 and 2016. The corpus is encoded in TEI. The corpus is available for online browsing through the Corpuscle concordancer (CLARINO distribution). For a relevant publication, see Meurer (2018).
Bulgarian National Reference Corpus (BNRC) **Size:** 70 million tokens	Bulgarian	This corpus[1] includes Bulgarian texts taken from news media, literature, and administrative documents between 1997 and 2002.

[1] There is another Bulgarian reference corpus (thanks to Petya Osenova for pointing this out), namely the 5.4 billion-word Bulgarian National Corpus (BNK) (Koeva et al. 2012) developed by the Institute for the Bulgarian Language, which however is not a member of the Bulgarian CLARIN consortium CLaDA-BG.

Annotation: tokenized, PoS-tagged **Licence:** Individual terms of agreement		The tokenised corpus is available through WebCLaRK (CLaDA-BG), while the PoS-tagged version is available only upon request. For a related publication, see Simov et al. (2004).
Croatian language corpus Riznica 0.1 **Size:** 101.8 million tokens, 85.3 million words, 4.7 million sentences, 14,781 texts **Annotation:** sentence segmented, PoS-tagged, lemmatized **Licence:** CC BY-NC-SA 4.0	Croatian	This corpus includes Croatian texts taken from fiction (28%) and specialised texts (72%). The corpus is available for online browsing via noSketch Engine and KonText and for download from the CLARIN.SI repository. For a related publication, see Ćavar and Brozović Rončević (2012).
Croatian National Corpus **Size:** 101 million tokens	Croatian	This corpus includes Croatian texts taken from newspapers, magazines, popular texts, and fiction. The corpus is available for online browsing through the noSketch Engine. For a relevant publication, see Tadić (2002).
SYN2005: balanced corpus of written Czech **Size:** 100 million words **Annotation:** MSD-tagged, lemmatized **Licence:** Czech National Corpus (Shuffled Corpus Data)	Czech	This corpus includes Czech texts published between 2000 and 2004. The corpus is encoded in XML. The corpus is available for online browsing through the KonText concordancer and can be downloaded from the LINDAT repository. For a relevant publication, see Hnátková et al. (2014).
SYN2010: balanced corpus of written Czech **Size:** 100 million words **Annotation:** MSD-tagged, lemmatized **Licence:** Czech National Corpus (Shuffled Corpus Data)	Czech	This corpus includes Czech fiction, professional literature, newspapers etc. published between 2005 and 2009. The corpus is encoded in XML. The corpus is available for online browsing through the KonText concordancer and can be downloaded from the LINDAT repository. For a relevant publication, see Hnátková et al. (2014).

SYN2015: representative corpus of written Czech **Size:** 100 million words **Annotation:** MSD-tagged, lemmatized **Licence:** Czech National Corpus (Shuffled Corpus Data)	Czech	This corpus includes Czech fiction, professional literature, newspapers etc. published between 2010 and 2014. The corpus is encoded in XML. The corpus is available for online browsing through the KonText concordancer and can be downloaded from the LINDAT repository. For a relevant publication, see Hnátková et al. (2014).
DK-CLARIN Reference Corpus of General Danish **Size:** 45.1 million words **Annotation:** PoS-tagged, sentence and paragraph segmentation, lemmatized **Licence:** CLARIN ACA-NC	Danish	This corpus includes Danish texts published between 2008 and 2011. The corpus is encoded in TEI. Non-linguistic metadata includes information on source and year of publication. The corpus is available for download from the CLARIN-DK repository.
SoNaR **Size:** 500 million words **Annotation:** PoS-tagged, lemmatized, named entities; coreference annotation and annotation of spatial and temporal relations for the manually annotated SoNaR-1 subset **Licence:** Terms of Agreement	Dutch	This corpus includes representative Dutch texts (fiction, brochures, magazines, legal texts, newspapers, parliamentary proceedings, and computer-mediated communication). Aside from written materials, the corpus also contains transcriptions of spoken language. The corpus is encoded in FoLiA. The corpus is available for online browsing through the OpenSONAR concordancer and can be downloaded from the Dutch Language Institute (CLARIAH-NL).
Corpus of Contemporary American English – Kielipankki version **Size:** 440 million words; 190,000 texts **Annotation:** PoS-tagged, lemmatized **Licence:** CLARIN ACA (online version); CLARIN RES (downloadable version)	English (American)	This corpus includes American English texts evenly divided into the spoken, fiction, magazine, newspaper, and academic genres (around 88 million words each) published between 1990 and 2012. The corpus is available for download from the Finnish Language Bank as well as for online browsing through the concordancer Korp (FIN-CLARIN distribution).
British National Corpus **Size:** 100 million words	English (British)	This corpus includes English texts (fiction, magazines, newspapers, and academic writing) published between 1980 and 1993.

Annotation: PoS-tagged, lemmatized **Licence:** BNC User Licence (restricted for the downloadable version)		The corpus is encoded in TEI. Non-linguistic metadata include contextual and bibliographic information. Aside from written materials, the corpus also includes transcriptions of spoken language. The corpus is available for online browsing through a dedicated concordancer and can be downloaded from the Oxford Text Archive (CLARIN-UK).
Estonian National Corpus 2019 **Size:** 1.5 billion words **Annotation:** MSD-tagged, lemmatized **Licence:** CC-BY-SA	Estonian	This corpus includes Estonian texts published between 1990 and 2019. Amongst others, this corpus contains the Estonian Reference Corpus as a subcorpus. The corpus is available for download from META-SHARE (CELR distribution).
Estonian Reference Corpus **Size:** 175 million words **Annotation:** MSD-tagged, lemmatized **Licence:** free for non-commercial use	Estonian	This corpus includes Estonian texts (fiction, PhD theses, newspapers, magazines, parliamentary transcriptions, computer-mediated communication) published between 1990 and 2007. The corpus is encoded in TEI. The corpus is available for online browsing through a dedicated concordancer and is available for download from CELR.
GNC: Georgian National Corpus **Size:** 217 million words **Annotation:** MSD-tagged, lemmatized **Licence:** CC-BY-NC, CLARIN_ACA-NC-LOC-PRIV-ND-*	Georgian (Old, Middle, Modern), Mingrelian, Svans	This corpus includes texts from languages spoken in Georgia from 500 to 2013. The corpus is encoded in TEI XML. The corpus is available for online browsing through a dedicated webpage. For a relevant publication, see Meurer (2017).
DeReKo **Size:** 31.7 billion words **Annotation:** MSD-tagged, lemmatized **Licence:** CC-BY-SA	German	This corpus includes German texts in a wide variety of genres published from 1947 onwards. Non-linguistic metadata include rich bibliographic information and partial layout information. Part of the corpus is available for download from a dedicated webpage (CLARIN-D distribution), while the entire corpus can be queried online through the COSMAS II platform.

		For a relevant publication, see Kupietz et al. (2018).
Corpus of Greek Texts **Size:** 27.6 million words **Annotation:** **Licence:** CC-BY-NC; ACA	Greek	This corpus includes representative Greek texts published between 1990 and 2010. Aside from written materials, the corpus also includes transcriptions of spoken language. The corpus is available for online browsing through a dedicated concordancer. For a relevant publication, see Goutsos (2010).
Diachronic corpus of Greek of the 20th century **Size:** 20 million words **Annotation:** **Licence:** CC BY-NC	Greek	This corpus includes Greek texts published in the 20th century. The corpus is available for download from CLARIN:EL.
Hellenic National Corpus **Size:** 47 million words **Annotation:** sentence segmented **Licence:** proprietary	Greek	This corpus includes Greek texts published from 1990 onwards. The corpus is available for online browsing through a dedicated concordancer. For a relevant publication, see Gavrilidou (2002).
Hungarian National Corpus **Size:** 190 million tokens **Annotation:** PoS-tagged	Hungarian	This corpus includes Hungarian texts (newspapers, literature, scientific articles, official and personal documents). The corpus is available for online browsing through a dedicated concordancer. For a relevant publication, see Váradi (2002).
The Icelandic Gigaword Corpus **Size:** 1.3 billion words **Annotation:** MSD-tagged, lemmatized **Licence:** CC-BY and a special user licence	Icelandic	This corpus includes Icelandic texts (newspapers, parliamentary proceedings, adjudications, fiction and non-fiction) published until 2017. The corpus is encoded in TEI. Non-linguistic metadata include bibliographic information. Aside from written materials, the corpus also contains transcriptions of spoken language. The corpus is available for online browsing and download through CLARIN-IS.

		For a relevant publication, see Steingrímsson et al. (2018).
Corpus of the Contemporary Lithuanian Language **Size:** 208.4 million tokens **Annotation:** MSD-tagged, lemmatized **Licence:** CLARIN RES	Lithuanian	This corpus includes Lithuanian texts (mostly newspapers but also fiction, non-fiction, and specialised magazines) published between 1990 and 2008. The corpus is encoded in TEI. Non-linguistic metadata includes bibliographic information. Aside from written materials, the corpus also contains transcriptions of spoken language. The corpus is available for online browsing through a dedicated concordancer.
The Lexicographic Corpus for Norwegian Bokmål (LBK) **Size:** 100 million tokens **Annotation:** PoS-tagged, lemmatized **Licence:** CLARIN_ACA-NC-LOC-ND	Norwegian (Bokmål)	This corpus includes representative Norwegian (Bokmål) texts (newspapers and periodicals, non-fiction, fiction, TV subtitles, and small print) published between 1985 and 2013. The corpus is available for online browsing through the concordancer Glossa (CLARINO). For a relevant publication, see Lain Knudsen and Vatvedt Fjeld (2013).
Norsk Ordboks Nynorskkorpus (NNK) **Size:** 107.8 million words **Annotation:** MSD-tagged, lemmatized **Licence:** CLARIN_RES-NC-DEP	Norwegian (Nynorsk)	This corpus includes representative Norwegian (Nynorsk) texts published between 1866 and 2012. The corpus is encoded in XML. The corpus is available for online browsing through the Corpuscle concordancer (CLARINO).
National Corpus of Polish **Size:** 1.8 billion tokens **Annotation:** MSD-tagged, lemmatized	Polish	This is a written and spoken corpus that includes representative Polish texts published between 1945 and 2010. The corpus is encoded in TEI. Non-linguistic metadata includes information on source, year of publication, text type, title, author. Aside from written materials, the corpus also includes transcriptions of spoken language. The corpus is available for online browsing through a dedicated concordancer. For the relevant publication, see Przepiórkowski et al. (2012).

PAROLE Portuguese Corpus **Size:** 3 million words **Annotation:** MSD-tagged, manually disambiguated **Licence:** ELRA	Portuguese	This corpus includes Portuguese texts (newspapers, books, periodicals, and miscellaneous texts) published between 1996 and 1997. The corpus is encoded in the PAROLE format. The corpus is available for download from the ELRA catalogue.
Written corpus ccGigafida 1.0 **Size:** 126.9 million tokens, 103.2 million words, 31,722 texts **Annotation:** MSD-tagged, lemmatized **Licence:** CC-BY-NC-SA 4.0	Slovenian	This corpus includes representative Slovenian texts (newspapers, magazines, computer-mediated communication, fiction and non-fiction) published between 1990 and 2011. The corpus is encoded in TEI. Non-linguistic metadata includes information on source, year of publication, text type, title, author. This corpus is a downloadable subset of the representative Gigafida corpus (version 1). It can be downloaded from the CLARIN.SI repository. For a relevant publication, see Erjavec and Logar (2012).
Written corpus ccKres 1.0 **Size:** 12.2 million tokens, 9.8 million words **Annotation:** MSD-tagged, lemmatized **Licence:** CC-BY	Slovenian	This corpus includes balanced Slovenian texts (newspapers, magazines, computer-mediated communication, fiction and non-fiction) published between 1990 and 2011. The corpus is encoded in TEI. Non-linguistic metadata includes information on source, year of publication, text type, title, author. This corpus is a downloadable subset of the balanced Kres corpus. It can be downloaded from the CLARIN.SI repository. For a relevant publication, see Erjavec and Logar (2012).
Written corpus Gigafida 2.0 **Size:** 1.3 billion tokens, 1.1 billion words, 38,310 texts **Annotation:** MSD-tagged, lemmatized **Licence:** Individual terms of agreement	Slovenian	This corpus includes representative Slovenian texts (newspapers, magazines, computer-mediated communication, fiction and non-fiction) published between 1990 and 2018. The corpus is encoded in TEI. Non-linguistic metadata includes information on source, year of publication, text type, title, author. The corpus is available for online browsing through the noSketch Engine concordancer

		(CLARIN.SI distribution), as well as through a dedicated search engine. For a relevant publication, see Krek et al. (2018).
Written corpus Kres 1.0 **Size:** 99 million words **Annotation:** MSD-tagged, lemmatized **Licence:** Individual terms of agreement	Slovenian	This corpus includes balanced Slovenian texts (newspapers, magazines, computer-mediated communication, fiction and non-fiction) published between 1990 and 2011. This corpus is a balanced subset of the representative Gigafida corpus (version 1). The corpus is encoded in TEI. Non-linguistic metadata includes information on source, year of publication, text type, title, author. The corpus is available for online browsing through a dedicated concordancer. For a relevant publication, see Krek et al. (2018).

2.1. The overview

2.1.1. Identification

All corpora, except the following 9 (31%) corpora can be found in the VLO. In the parentheses, we list the CLARIN consortium with which the corpus is associated.

(1) AbNC: Abkhaz National Corpus (CLARINO)
(2) GNC: Georgian National Corpus (CLARINO)
(3) Norsk Ordboks Nynorskkorpus (NNK) (CLARINO)
(4) Hellenic National Corpus (CLARIN:EL)
(5) Corpus of Greek Texts (CLARIN:EL)
(6) Diachronic corpus of Greek of the 20th century (CLARIN:EL)
(7) Written corpus Gigafida 2.0 (CLARIN.SI)
(8) Written corpus Kres 1.0 (CLARIN.SI)
(9) Bulgarian National Reference Corpus (BNRC) (CLaDA-BG)

The CLARIN:EL corpora in (4)–(6) cannot be found in the VLO because the CLARIN:EL Central Inventory is not yet harvested, while the CLARIN.SI corpora in (7)–(8) are not listed in the VLO because they do not have CLARIN.SI repository entries. The CLaDA-BG corpus in (9) is presumably not listed in the VLO because CLaDA-BG has not yet established a B-Certified repostiry. It is unclear why the CLARINO corpora in (1)–(3) are not listed in the VLO.

2.1.2. Availability

2.1.2.1. For download and online querying

The following 10 (34%) corpora are available for download and for online querying. In the parentheses, we specify the repository from which the corpus can be downloaded and the concordancer.

(1) SYN2015: representative corpus of written Czech (LINDAT, KonText)
(2) SYN2010: balanced corpus of written Czech (LINDAT, KonText)
(3) SYN2005: balanced corpus of written Czech (LINDAT, Kontext)
(4) Corpus of Contemporary American English – Kielipankki version (FIN-CLARIN, Korp)
(5) SoNaR (CLARIAH-NL, OpenSONAR)
(6) British National Corpus (CLARIN-UK, the English-corpora concordancer)
(7) Estonian Reference Corpus (LINDAT/CELR, the Keeleveeb concordancer)
(8) DeReKo (CLARIN-D, COCMAS II)
(9) The Icelandic Gigaword Corpus (CLARIN.IS, dedicated concordancer)
(10) Croatian language corpus Riznica 0.1 (CLARIN.SI, noSketchEngine and KonText)

It should be noted that all the corpora are available for download in their entirety except the German reference corpus DeReKo, in which case only a few subcorpora can be downloaded under various licence agreements (see here) but not the entire corpus. Furthermore, 5 corpora (SYN2015: representative corpus of written Czech, SYN2010: balanced corpus of written Czech, SYN2005: balanced corpus of written Czech, SoNaR, and British National Corpus) can be downloaded directly from their respective repository, while Estonian Reference Corpus and The Icelandic Gigaword Corpus are only listed in the LINDAT and CLARIN-IS repositories, respectively, but are downloadable from their own dedicated webpages. At least in the case of The Icelandic Gigaword Corpus, this is likely due to the fact that the corpus is available for download in two subsets, each of which has its own licence.

2.1.2.2. For online querying

The following 13 (45%) corpora are available only for online querying. In the parentheses, we specify the repository with which the corpus is associated and the concordancer.

(1) AbNC: Abkhaz National Corpus (CLARINO, dedicated)
(2) Hellenic National Corpus (CLARIN:EL, dedicated)
(3) Corpus of Greek Texts (CLARIN:EL, dedicated)
(4) Croatian National Corpus (LINDAT, noSketch Engine)
(5) Hungarian National Corpus (LINDAT and HUN-CLARIN, dedicated)
(6) Corpus of the Contemporary Lithuanian Language (CLARIN-LT, dedicated)
(7) GNC: Georgian National Corpus (CLARINO, dedicated)
(8) The Lexicographic Corpus for Norwegian Bokmål (LBK) (CLARINO, Glossa)
(9) Norsk Ordboks Nynorskkorpus (NNK) (CLARINO, Corpuscle)
(10) National Corpus of Polish (LINDAT and CLARIN-PL, dedicated)
(11) Written corpus Gigafida 2.0 (CLARIN.SI, noSketch Engine and dedicated)
(12) Written corpus Kres 1.0 (CLARIN.SI, dedicated)
(13) Bulgarian National Reference Corpus (BNRC) (CLaDA-BG, WebCLaRK)

Note that the Croatian National Corpus cannot be accessed directly through LINDAT because the link to the external landing page (http://hnk.ffzg.hr/) is broken.

2.1.2.3. For download

The following 5 (17%) corpora are available only for download. In the parentheses, we specify the repository from which the corpus can be downloaded.

(1) DK-CLARIN Reference Corpus of General Danish (CLARIN-DK)

(2) Estonian National Corpus 2019 (CELR through META-SHARE)
(3) PAROLE Portuguese Corpus (PORTULAN through ELRA)
(4) Written corpus ccGigafida 1.0 (CLARIN.SI)
(5) Written corpus ccKres 1.0 (CLARIN.SI)

2.1.2.4. Unavailable

The following 1 (3%) corpus is not available. Note that it is unclear from the description of the corpus in CLARIN:EL whether it is still in development or not.

(1) Diachronic corpus of Greek of the 20th century (CLARIN:EL)

2.1.3. Metadata

2.1.3.1. Language

As is excepted for reference corpora, almost all corpora, i.e., 28 (97%) out of 29 – are monolingual, accounting for the following 17 languages:

(1) Slovenian (4 corpora)
(2) Czech (3 corpora)
(3) Greek (3 corpora)
(4) Croatian (2 corpora)
(5) English (2 corpora)
(6) Estonian (2 corpora)
(7) Norwegian (2 corpora)
(8) Abkhaz (1 corpus)
(9) Bulgarian (1 corpus)
(10) Danish (1 corpus)
(11) Dutch (1 corpus)
(12) German (1 corpus)
(13) Hungarian (1 corpus)
(14) Icelandic (1 corpus)
(15) Lithuanian (1 corpus)
(16) Polish (1 corpus)
(17) Portuguese (1 corpus)

An exception is the multilingual corpus (GNC: Georgian National Corpus), which covers languages spoken in Georgia – that is, Georgian (Modern, Middle, and Old), Mingrelian, and the Svans language.

2.1.3.2. Size and time period

All corpora contain information on size.

The size of the corpora can be summarised as follows:

- 1 very small corpora (<10 million words/tokens)
- 8 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)

The largest corpus is the German reference corpus DeReKo with 31.7 billion words, while the smallest corpus is PAROLE Portuguese Corpus with 3 million words.

16 (55%) of the corpora contain materials published from 1990 onward. In parentheses, we specify the time coverage.

(1) SYN2015: representative corpus of written Czech (2010–2014)
(2) SYN2010: balanced corpus of written Czech (2005–2009)
(3) SYN2005: balanced corpus of written Czech (2000–2004)
(4) Bulgarian National Reference Corpus (BNRC) (1997–2002)
(5) DK-CLARIN Reference Corpus of General Danish (2008–2011)
(6) Estonian Reference Corpus (1990–2007)
(7) Estonian National Corpus 2019 (1990–2019)
(8) Hellenic National Corpus (1990–)
(9) Corpus of Greek Texts (1990–2010)
(10) Corpus of Contemporary American English – Kielipankki version (1990–2012)
(11) Corpus of the Contemporary Lithuanian Language (1990–2008)
(12) PAROLE Portuguese Corpus (1996–1997)
(13) Written corpus ccGigafida 1.0 (1990–2011)
(14) Written corpus ccKres 1.0 (1990–2011)
(15) Written corpus Gigafida 2.0 (1990–2018)
(16) Written corpus Kres 1.0 (1990–2011)

8 (28%) corpora also cover older materials, as follows:

(1) DeReKo (1947–)
(2) AbNC: Abkhaz National Corpus (1920–2016)
(3) GNC: Georgian National Corpus (500–2013)
(4) The Lexicographic Corpus for Norwegian Bokmål (LBK) (1985–2013)
(5) Norsk Ordboks Nynorskkorpus (NNK) (1866–2012)
(6) National Corpus of Polish (1945–2010)
(7) Diachronic corpus of Greek of the 20th century (the 20th century)
(8) British National Corpus (1980–1993)

The following 4 (14%) corpora do not specify the time span of the materials.

(1) Croatian National Corpus
(2) SoNaR
(3) Hungarian National Corpus
(4) Croatian language corpus Riznica 0.1

Lastly, The Icelandic Gigaword Corpus states only the endpoint of the time coverage (i.e., materials published before 2017).

2.1.3.3. Annotation

The vast majority of the reference corpora (26 or 90% out of 29) specify the levels of linguistic annotation. These are as follows:

- MSD/PoS tagging (24 corpora)

- Lemmatization (22 corpora)
- Sentence segmentation (2 corpora)[2]
- Named entity recognition and coreference marking (1 corpus – i.e., SoNaR)

The following 3 (10%) corpora do not contain information on linguistic annotation:

(1) Croatian National Corpus
(2) Corpus of Greek Texts
(3) Diachronic corpus of Greek of the 20th century

2.1.3.4. Licence

26 out of 29 (90%) corpora contain information on licence. The following corpora lack this information:

(1) Croatian National Corpus
(2) Hungarian National Corpus
(3) National Corpus of Polish

In comparison with the rest of the resource families, the licences are quite varied, with only 8 corpora being available under CC-BY, 5 under CLARIN ACA, and 3 under CLARIN RES, while the rest of the corpora are available under various proprietary and individual licences.

It should also be noted that sometimes, the licences are known but are listed with the "unspecified availability" tag in the VLO. This is the case of the following corpus, which has a restricted licence.

(1) Corpus of the Contemporary Lithuanian Language (CLARIN RES)

The licence tag is unknown in the VLO because the relevant repository entry does not specify the licence, given that the corpus is not deposited there but is accessible from a dedicated webpage.

3. Conclusion

In this report, we gave an overview of 29 reference corpora in the CLARIN infrastructure. We presented their identification (i.e., whether they have VLO entries) and their availability (for download, online querying or both), as well as 5 types of metadata – language, size, publication period, annotation, and licence.

In terms of identification, 9 (31%) out of the 29 identified corpora are not listed in the VLO. In terms of availability, 1 (3%) out of the 29 corpora is unavailable for download and online querying. Otherwise, availability is as follows: 10 (34%) corpora are available both for online querying and download, 5 (17%) corpora are available only for download, and 13 (45%) corpora are available only for online querying. It is noteworthy all of the reference corpora can be downloaded in their entirety save for the German reference corpus, in which case only a few subcorpora can be downloaded under various licence agreements.

[2] Presumably, more corpora are sentence segmented, although they do not explicitly specify this information in their documentation.

In terms of language, almost all (28 or 97%) corpora are monolingual. All corpora cover languages spoken in Europe, with Slovenian, Czech, and Greek being the most represented languages. The only multilingual corpus contains the national languages spoken in Georgia.

Information on size is available for all the 29 corpora. It is noteworthy that 6 (21%) corpora contain more than 1 billion words/tokens, while only 1 (3%) corpus contains fewer than 10 million words/tokens. Most (14 or 48%) corpora contain between 100 million and 1 billion words/tokens. Information on time period is generally well included, and is not specified for only 4 (14%) corpora. 16 (55%) of the corpora contain materials published from 1990 onward.

Information on annotation fares the best out of all the surveyed resource families, and is available for 26 (90%) corpora, with 24 (83%) corpora being MSD/PoS-tagged. Licence is also readily included and is available for 26 corpora, though it must be noted that the licence is listed as unknown in the case of certain corpora whose licence is otherwise known but not specified in the repository harvested by the VLO.

Lastly, it should be noted that the documentation of the genres constituting the reference corpora is, by and large, under-informative in the CLARIN repositories. The repositories either completely omit this information or only exemplify some of the included genres (e.g., “this corpus contains a wide variety of written genres, such as newspapers, fiction, etc.”), while the full break-down including the balance of the genres is only available on the external project pages. Given that genres are arguably one of the key pieces of metadata in regard to reference corpora, we suggest that they be rigorously documented in the repositories as well.

4. References

Ćavar, Damir, and Dunja Brozović Rončević. 2012. Riznica: the Croatian Language Corpus. *Prace filologiczne*, 63: 51–65. http://riznica.ihjj.hr/CLC-Slavicorp.pdf.

Erjavec, Tomaž, and Nataša Logar Berginc. Referenčni korpusi slovenskega jezika (cc)Gigafida in (cc)KRES. 2012 In *Zbornik Osme konference Jezikovne tehnologije*, 57–62. http://nl.ijs.si/isjt12/proceedings/isjt2012_11.pdf.

Gavrilidou, Maria. 2002. The Hellenic National Corpus on-line. *Revue belge de Philologie et d'Histoire*, 80 (3): 1003–1015. https://www.persee.fr/doc/rbph_0035-0818_2002_num_80_3_4652.

Goutsos, Dionysis. 2010. The Corpus of Greek Texts: a reference corpus for Modern Greek. *Corpora*, 5 (1): 29–44. https://doi.org/10.3366/E1749503210000353.

Hnátková, Milena, Michal Kren, Pavel Procházka, and Hana Skoumalová. 2014. The SYN-series corpora of written Czech. In *Proceedings of LREC 2014*, 160–164. http://www.lrec-conf.org/proceedings/lrec2014/pdf/294_Paper.pdf.

Koeva, Svetla, Ivelina Stoyanova, Svetlozara Leseva, Tsvetana Dimitrova, Rositsa Dekova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 0 (1): 65–110. http://dx.doi.org/10.15398/jlm.v0i1.33.

Krek, Simon, Polona Gantar, Špela Arhar Holdt, and Vojko Gorjanc. 2016. In *Proceedings of the Conference on Language Technologies and Digital Humanities*, 200–202. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Krek-et-al_Nadgradnja-korpusov-Gigafida-Kres-ccGigafida-ccKres.pdf.

Kupietz, Marc, Harald Lüngen, Pawel Kamocki, and Andreas Witt. 2018. In *Proceedings of LREC 2018*, 4353–4360. http://www.lrec-conf.org/proceedings/lrec2018/summaries/737.html.

Lain Knudsen, Rune, and Ruth Vatvedt Fjeld. 2013. LBK2013: A balanced; annotated national corpus for Norwegian Bokmål. In *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013*, 12–20. https://ep.liu.se/ecp/article.asp?issue=088&article=003&volume=0.

Leech, Geoffrey. 2002. The Importance of Reference Corpora. https://www.uzei.eus/wp-content/uploads/2017/06/06-Geoffrey-LEECH.pdf.

Meurer, Paul. 2017. The Morphosyntactic Analysis of Georgian. http://clarino.uib.no/gnc/doc/Morphosyntactic-analysis-of-Georgian.pdf.

Meurer, Paul. 2018. The Abkhaz National Corpus. In *Proceedings LREC 2018*, 2456–2460. http://www.lrec-conf.org/proceedings/lrec2018/pdf/548.pdf.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. *Narodowy Korpus Języka Polskiego*. http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf.

Simov, Kiril, Petya Osenova, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff. 2004. A Language Resources Infrastructure for Bulgarian. In *Proceedings of LREC 2004*, 1685–1688. http://www.lrec-conf.org/proceedings/lrec2004/summaries/316.html.

Steingrímsson, Steinþór, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of LREC 2018*, 4361–4366. https://www.aclweb.org/anthology/L18-1690.

Tadić, Marko. 2002. Building the Croatian National Corpus. In *Proceedings of LREC 2002*, 441–446. http://www.lrec-conf.org/proceedings/lrec2002/pdf/170.pdf.

Váradi, Tamás. 2002. The Hungarian National Corpus. In *Proceedings of LREC 2002*, 385–389. http://www.lrec-conf.org/proceedings/lrec2002/pdf/217.pdf.