

The CLARIN Resource Families Workflow

Jakob Lenardič

Institute of Contemporary History
Ljubljana, Slovenia
jakob.lenardic@inz.si

Alexander König

CLARIN ERIC
alex@clarin.eu

Kristina Pahor de Maiti Tekavčič

Institute of Contemporary History
Ljubljana, Slovenia &
Faculty of Arts, University of Ljubljana
kristina.pahordemaiti@ff.uni-lj.si

Dieter Van Uytvanck

CLARIN ERIC
dieter@clarin.eu

Abstract

This paper presents the new workflow for the CLARIN Resource Families (CRF) initiative. After introducing the CRF initiative and laying out its primary aim, we present the functioning of the new CRF backend, which uses a JSON-based workflow publicly accessible on GitHub. We then present the guidelines for working with the JSON files and describing the metadata. We also discuss how the new workflow enables a collaborative approach to maintaining the families. Finally, we discuss how the CRF complements and is complemented by CLARIN ERIC's Virtual Language Observatory.

1 Introduction

A key aim of the CLARIN Resource Families (CRF) initiative (Lenardič & Fišer, 2022b) is the promotion of mature language resources and technologies (LRTs), such as language corpora, lexical resources, and tools, that are provided by national CLARIN consortia. The way in which the CRF initiative specifically contributes towards this goal is by offering access to manually curated overviews of LRTs, which are grouped together along thematic lines into 'families', such as *parliamentary corpora*, *tools of sentiment analysis*, *language models*, and so on. Importantly, all other key CLARIN services that automatically collate LRTs dispersed throughout the CLARIN infrastructure, such as the Virtual Language Observatory (VLO; Goosen and Eckart, 2014; Van Uytvanck et al., 2010, 2012), do not provide search facets that would allow researchers to pick out tools or resources belonging just to a particular family to the exclusion of everything else. The lack of such search options also follows in part from the fact that taxonomic categories like parliamentary corpora do not correspond to metadata components in CMDI (Windhouwer & Goosen, 2022). By providing the manually curated thematic groupings, the CRF initiative fills this gap in resource and tool discovery, and thereby facilitates the increased findability of CLARIN LRTs from the perspective of the FAIR principles (Wilkinson et al., 2016).

When the CRF initiative launched in 2018, it consisted of just 4 corpus families, whose creation had a project-internal motivation at the time (Fišer et al., 2018).¹ However, the initiative soon proved popular especially amongst non-CLARIN-affiliated researchers in the (digital) humanities and social sciences, who often first encounter the CLARIN infrastructure through a CRF webpage.² Consequently, the CRF overviews have been iteratively expanded throughout the years with the addition of new families and LRTs so that on 1 April 2026 there are 16 corpus families, 6 families of lexical resources, and 6 tool families, which together amount to around 1600 manually curated LRTs.

Because of this expansion, there is a growing danger that the individual families will sooner or later become deprecated, especially since the upkeep of the families has thus far been largely a CLARIN ERIC-internal affair. We have therefore developed a new backend for the CRF overviews that is more

¹Specifically, it was tied to workshops organised in the context of the CLARIN Plus project: <https://www.clarin.eu/content/factsheet-clarin-plus>.

²CRF webpages are typically the first or second result on Google for search terms like 'parliamentary corpora' or 'parallel corpora'.

geared towards collaboration than what was previously the case. In this paper, we present the functioning of this new backend and the ways in which involvement has now become possible.

The rest of the paper is organised as follows. In Section 2, we first introduce the technical aspects of the new workflow and guidelines for using it. In Section 3, we present the ways in which collaborators can become involved in the upkeep of the CRF overviews as well as our own ongoing efforts for facilitating a greater degree of external involvement. In Section 4, we discuss how the CRF complements and is in turn complemented by the VLO both technologically and conceptually. Section 5 concludes the paper.

2 The GitHub-based workflow for creating CRF overviews

2.1 Main characteristics

```

1 {
2   "Name": "Slovenian parliamentary corpus siParl 4.0",
3   "URL": "http://hdl.handle.net/11356/1936",
4   "Family": "Parliamentary corpora",
5   "Description": "The corpus contains Slovenian parliamentary debates from 1990 to
6     2022.\n\nThe corpus is encoded according to the <a href=\"
7     https://github.com/clarin-eric/parla-clarin/\">Parla-CLARIN</a> schema. In terms
8     of linguistic annotation, the part-of-speech tags, morphological analyses,
9     and syntactic parses in the corpus follow the <a href=\"
10    https://universaldependencies.org/sl/\">Universal Dependencies</a> framework.\n
11    The corpus is available for download from the CLARIN.SI repository and for
12    online querying through the noSketch Engine concordancer.",
13   "Language": ["slv"],
14   "Licence": "CC BY 4.0",
15   "Size": ["13,233 texts", "1,223,250 utterances", "239,224,540 words",
16     "273,023,945 tokens"],
17   "Annotation": ["tokenised", "PoS-tagged", "lemmatised", "syntactically parsed",
18     "annotated for named entities"],
19   "Infrastructure": "CLARIN",
20   "Access": {
21     "Concordancer": "https://www.clarin.si/ske/#dashboard?corpname=siparl40",
22     "Download": "http://hdl.handle.net/11356/1936"
23   },
24   "Publication": "https://www.zotero.org/groups/562080/clarin/items/K24SGCXC"
25 }
```

Figure 1: A CRF JSON file.

The new CRF backend is using a Python-based workflow on GitHub.³ The workflow takes as input a set of JSON files located in three folders (*corpora*, *lexical-resources* and *tools*) and then, with the help of Drupal, converts them to tabular HTML code that is periodically and automatically fed into the online overviews hosted at <https://www.clarin.eu/resource-families>. Each JSON file corresponds to an entry in the tabular online overviews.

The built-in Drupal features ensure that the CRF overviews have started behaving in a similar way to a generic search engine. The overviews are therefore no longer entirely static – it has become for instance possible to filter the CRF entries so that only resources in a specific language or with a specific annotation layer are shown. It is also possible to use a single filter across all of the CRF overviews to generate a list of high-quality resources for a specific language or even annotation layer, which is a feature that CRF users have long been asking for.

For instance, the workflow maps the contents of the JSON file in Figure 1, which is for the Slovenian parliamentary corpus *siParl 4.0* (Pančur et al., 2024), to HTML code that is then included as part of the Parliamentary corpora resource family.⁴ The organisation into subsections seen on the CRF webpages, such as the distinction between CLARIN-affiliated and CLARIN-external LRTs (a distinction that is being phased out; see Section 2.3), along with the division between monolingual or multilingual LRTs, is derived

³<https://github.com/clarin-eric/clarin-resource-families/>

⁴<https://www.clarin.eu/resource-families/parliamentary-corpora>

from name/value pairs such as the one for *Infrastructure* (line 10 in the figure; here concretely specified as CLARIN-internal) and the number of values in the *Language* array (specified just for Slovenian in line 6; hence, the JSON CRF entry for *siParl 4.0* is mapped to the Monolingual subgroup).

The name/value pairs shown in the figure mostly recur throughout the corpus and lexical-resource families, but not the tool families, so potential users of the workflow are instructed to re-use an existing file in the repository whenever they begin to work on a particular family.

2.2 The guidelines for the workflow

For working with the GitHub-based workflow, a set of guidelines has been prepared by Lenardič and König (2025). The guidelines are based on the existing recommendations of the CLARIN.SI consortium for the documentation of language resources,⁵ and on proposals by Odijk (2019, pp. 122–123) for the documentation of language-processing tools.

In the first part, the guidelines present a beginner-friendly tutorial for editing and creating the JSON files. This tutorial assumes no prior knowledge of JSON syntax, so it also highlights rather trivial but often recurring JSON requirements, such as the obligatory use of `\n` operators for line or paragraph breaks and the usage and function of quotation marks.⁶

The second part of the guidelines, however, is more descriptive in nature, and is crucial for researchers both with or without experience in working with JSON syntax. This part lays out the requirements for the JSON values, some of which have been semi-standardised following recommendations in the related literature (e.g., Bański et al., 2025).

In what follows, we discuss the guidelines for each name–value pair in turn while highlighting some existing non-trivial issues as well.

2.2.1 Language codes

The values for the *Language* array in line 6 in Figure 1 must be specified using the ISO 639-3 language set, which is in line with the recommendation of the CLARIN Standards and Interoperability Committee (Bański & Hedeland, 2022).

In terms of natural-language coverage, the ISO 639 set is quite comprehensive, specifying three-letter codes for ‘spoken, written and signed languages, including extinct or historical languages’ (Romary, 2025, p. 40). For instance, extinct languages in CRF include corpora in Latin and Old Norse, such as the *LatinISE corpus (version 5)* (McGillivray, 2025) and *The Saga Corpus (Fornritin)* (Helgadóttir & Barkarson, 2018), whose languages are specified with the ISO 639-3 codes *lat* and *non* respectively.

We note, though, that the ISO 639 codes sometimes do not provide the greatest coverage when it comes to varieties within a national language, even those widely spoken in Europe and the Americas. For instance, the Portuguese CLARIN node PORTULAN hosts several corpora dedicated to variants of Portuguese spoken not only in Portugal, but across the world as well. One such resource is *CorPop: a corpus of popular Brazilian Portuguese* (Finatto & Grilo, 2020). However, there is no ISO 639-3 code that would distinguish Brazilian Portuguese from its European variant.⁷

There is a language catalogue whose granularity is finer than that of ISO 639; namely, Glottolog (Nordhoff, 2012).⁸ If CRF were to adopt Glottolog for reasons of better language coverage, then *CorPop: a corpus of popular Brazilian Portuguese* (Finatto & Grilo, 2020) could employ two of the so-called glottocodes (see Forkel & Hammarström, 2022) – namely, *port1283* for the Portuguese macro variant,⁹ whose corresponding ISO 639-3 code is *por*,¹⁰ as well as *braz1246* for the Brazilian subvariety,¹¹ which

⁵<https://www.clarin.si/repository/xmlui/page/deposit>

⁶The paragraph breaks in the CRF descriptions do not necessarily correspond to the breaks in the ‘free-text’ description in the original metadata. Sometimes, the latter descriptions do not use paragraph breaks at all, or any other textual formatting either, in contrast to the CRF ones, which strive to be consistent about text placement and layout. Note that CRF descriptions are not meant to be copies of the descriptions in the repository entries, but rather serve as partially unified descriptions briefly recapping the main research-domain-specific characteristics of the data.

⁷<https://tinyurl.com/ezdxh9bj>

⁸We thank Joshua Wilbur for bringing this database to our attention.

⁹<http://glottolog.org/resource/linguoid/id/port1283>

¹⁰<https://iso639-3.sil.org/code/por>

¹¹<http://glottolog.org/resource/linguoid/id/braz1246>

lacks the corresponding ISO 639-3 code.

Another macro language that could benefit from a move to Glottolog is English, which is still the language of the majority of CRF corpora. In the 5.3 version of Glottolog (Hammarström et al., 2026), there are 34 glottocodes for variants of English under ‘Macro English’,¹² whereas there are 23 ISO 639-3 codes for the same macro language.¹³ For the purposes of the CRF initiative, a switch to Glottolog would then mean that a corpus like *Corpus of Historical American English – Kielipankki Korp version 2017H1* (Davies, 2017), which belongs to the Historical corpora family,¹⁴ could be tagged with the glottocode *nort3314*,¹⁵ which denotes the North American variant and does not have an ISO 639-3 equivalent.

While the coverage of Glottolog is greater than that of ISO 639, there is a downside from the perspective of manual curation, which is the complexity of the glottocodes themselves; compare the ISO 639-3 code for Portuguese – that is, *por* – with the corresponding and lengthier glottocode – that is, *port1283*. This could be problematic in case curators type out the JSON values by hand instead of copying them from somewhere, which they often do in fact. Thus, CRF is for the time being sticking with the ISO 639-3 set, although a switch to Glottolog might be considered in the future due to its greater breadth of linguistic coverage.

2.2.2 Licence and size

Licence (line 7 in Figure 1) is written using the canonical abbreviation (i.e., CC BY 4.0 in this case), although in the future another name–value pair might be added for licence URL, which would be especially helpful for lesser known licences such as CLARIN PUB.¹⁶

It is worth mentioning, however, that not all CLARIN B- or C-centres consistently use the canonical abbreviations – for instance, the licence of the lexical resource *ViPER verb lexical database* (Mamede et al., 2020) is spelt as ‘CC - BY - NC - CD’ rather than ‘CC BY-NC-CD’ (notice also the missing licence version) in the original metadata records. Even in such cases, the CRF overviews stick to the canonical forms, which is in line with the idea from Section 2.2.4 and Footnote 6 that CRF entries are not meant to be copies of metadata records from the repositories, especially if there are issues with the latter.

Size (line 8) is provided in as many sensible categories as possible, where the categories should be relevant for cross-resource comparison (e.g., number of tokens vs. number of words). Just like in the case of licence, the CRF overviews deviate from the metadata records when the latter provide size only in bytes or the number of files included in the deposit, as such information – although common – is not really valuable from a comparative perspective.

2.2.3 Annotation

Annotation (line 9 in Figure 1) uses past-participle forms, so that linguistic annotation layers are described with terms like *tokenised*, *lemmatised*, *PoS-tagged*, and *syntactically parsed*. Sometimes, we use *MSD-tagged* for part-of-speech tags that define complex morphosyntactic features,¹⁷ as for instance those used by Slavic corpora.

An issue with this past-participle approach, however, is that certain other annotation processes such as ‘named entity recognition’ do not trivially admit such forms. The reason why we are currently sticking to the flawed past-participle approach is that there does not seem to exist a consolidated vocabulary of annotation layers that could be re-used in the overviews; see Lenardič (2023) for further discussion about the diversity of annotation descriptors within CLARIN. We note that a possible solution would be to switch to simple nominal forms, so that the Annotation layers could simply be listed with descriptors like ‘tokens’, ‘lemmas’, and ‘PoS tags’ instead of the past participles.

Importantly, the *Annotation* JSON name–value pair is meant as an array for listing the annotation layers, not for describing them. For further specifying any additional characteristics of the annotation

¹²<https://glottolog.org/resource/languoid/id/stan1293>

¹³<https://tinyurl.com/rfr4vrnp>

¹⁴<https://www.clarin.eu/resource-families/historical-corpora>

¹⁵<https://glottolog.org/resource/languoid/id/nort3314>

¹⁶<https://www.clarin.eu/content/licenses-and-clarin-categories#1-pub-language-resources>

¹⁷The acronym MSD refers to ‘a morphosyntactic description of a token, such as the Universal Dependencies morphological features’ (Ljubešić & Erjavec, 2025, p. 81).

process, curators are encouraged to use the *Description* name–value pair in line 5 of the figure, to which we know turn.

2.2.4 The ‘free-text’ description

A problem that commonly recurs across the CLARIN infrastructure is that the deposited LRTs are described in various under-informative ways. Consequently, it is often difficult to determine what kind of data the LRT actually consists of or what the functionality is in the case of tools, especially if the name of the LRT in question is opaque (see Lenardič & Fišer, 2022a, for concrete examples).

The guidelines therefore pay particular attention to the value of the *Description* field (line 5 in Figure 1), providing a list of suggestions for describing each type of LRT separately. We now overview these recommendations for each type in turn.

For describing corpora, curators are advised to provide at least some of the following information: modality (spoken, written, visual, etc.), time period (based on publication date), geographic coverage, data sampling (text types and their ratios; text sources and their ratios), envisaged research domains, and any important and unique domain-specific characteristics (e.g., the range of participants’ ages and L1s in L2 learner corpora).¹⁸ Curators are also encouraged to use this name–value pair for further specifying the characteristics of linguistic annotation, including which morphosyntactic tagset (e.g., MULTEXT-East; Erjavec, 2012, 2017) or syntactic framework (e.g., Universal Dependencies; De Marneffe et al., 2021) has been used for tagging or parsing the corpus.

For lexical resources, curators are advised to tailor the description to the subtype of the resource. For training lexica (e.g., *MorfFlex CZ 2.1*; Hajič et al., 2024), they should describe the key features of the annotation process (e.g., tagset, manual mark-up) and other structural information (e.g., lemma frequencies). For dictionaries (e.g., *Words of the 16th-Century Slovenian Literary Language*; Ahačič et al., 2011), curators should briefly overview how the entries are structured and what kind of grammatical (e.g., morphological information, collocation properties, phonemic transcription) and non-grammatical information (examples of use, contextual features) is included.

Lastly, for software, curators are encouraged to provide information on the tool’s applicability in terms of the relevant research domain(s), its distribution and installation requirements, as well as its output and input characteristics, which includes MIME-types, annotation schemata, and tagsets. Furthermore, they should provide information that is unique to the functionality, such as the types and granularity of categories recognised by a named entity recogniser (e.g., *NameTag*; Straka and Straková, 2014; Straková and Straka, 2025) or the types (e.g., sentence-level, document-level) and levels (binary, ternary, quaternary, etc.) of sentiment recognised by a sentiment analyser (e.g., *OptaHopper*; Žak and Skuczyńska, 2018).

The following is a CRF description of the Slovenian *Gigafida 2.0* (Krek et al., 2019) reference corpus. The first paragraph provides brief information about the corpus by summarising the modality, data sources and time period; the second paragraph specifies encoding and non-linguistic annotation; lastly, the third paragraph specifies the availability.

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 include information on source, year of publication, text type, title, and author.

The corpus is available for online browsing through the noSketch Engine concordancer (CLARIN.SI distribution), as well as through a dedicated search engine.

What is important is that the description is not the same as that of the original metadata record (see Krek et al., 2019), as the latter is quite lengthier in fact. These sort of succinct three-to-four paragraph descriptions are consistently used in CRF overviews. Ensuring such presentational consistency is particularly important now that external collaborators have become involved in the initiative (Section 3).

¹⁸<https://www.clarin.eu/resource-families/L2-corpora>

2.2.5 Publication

Publication values (line 15 in Figure 1) correspond to hyperlinks to records within the CLARIN Resource Families subfolder of the CLARIN Zotero Library.¹⁹ In the figure, the hyperlink is to the Zotero record of the paper by Meden et al. (2025), which describes the *siParl 4.0* (Pančur et al., 2024) corpus.

2.3 Inclusion criteria

CRF stopped accepting submissions of CLARIN-external LRTs in 2025; however, the existing ‘CLARIN external’ subsections are for now being kept as legacy pages. If these are removed at one point (though the JSONs will not be lost due to GitHub’s Version Control), then the ‘Infrastructure’ name–value pair (line 10 in Figure 1) might acquire another function, perhaps to specify the CLARIN centre associated with the LRT.

We now briefly address what is ‘internal’ to CLARIN and can therefore be considered as a candidate for inclusion in the CRF overviews. The main criterion for inclusion is that corpora and lexical resources must be deposited in a CLARIN C- or B-centre. Such resources are consequently also findable through the VLO. This criterion is less stringent for tools, which are often found in repositories that are better tailored to software, such as GitHub.

Importantly, CRF also no longer hosts resources or tools that are wholly inaccessible – that is, a resource or tool must either be deposited with a C- or B-centre or be accessible online; in Figure 1, for instance, the online accessibility of *siParl 4.0* (Pančur et al., 2024) is concretely defined as being through CLARIN.SI’s noSketch Engine concordancer. In other words, this means that the *Access* array beginning in line 11 of Figure 1 should not be empty. The motivation behind this decision is tied to CRF’s mission to provide listings of *reusable and mature* LRTs important for research within the (digital) humanities and social sciences. This is also a point in which the CRF overviews differ from the VLO (see Section 4), as the latter also provides listings of LRTs that are only accompanied by the metadata, but do not include the data themselves.

Another difference from the VLO is that there should generally be a consistent level of granularity in the CRF – only entire resources (e.g., an entire corpus) are typically included in CRF, but not their possible subsets, which sometimes correspond to their own repository entries. Similarly, only one (i.e., typically the latest one) version of a resource or tool is included in a CRF overview.

2.4 Metadata curation

As already stated, CRF entries are rarely just repeats of the metadata records in the CLARIN repositories; aside from producing concise descriptions comparable across the resources and tools (Section 2.2.4), they try to fill in potential gaps in metadata. The latter happens whenever any metadata are missing from the CMDI record of the deposited entry but turn out to be actually available elsewhere – typically in presentational materials related to the LRT or directly from the author(s).

As a by-product of this, the CRF initiative also contributes to ongoing curation efforts at the CLARIN ERIC level. This is done by maintaining a list of GitHub issues that spell out all the missing or unclear metadata identified whenever a new LRT is added to the CRF overviews or even when a new family is created. These GitHub tickets are assigned via the labelling system to national consortia whose representatives are then expected to resolve them.²⁰ While these issues were until recently collated in a dedicated repository,²¹ they are now going to be listed as part of the new workflow (Section 2.1) and linked to the JSON files themselves.

Centre	CRF Family
CKCMC	CMC Corpora
ACE	Corpora of Disordered Speech
CKL2CORPORA	L2 Learner Corpora
CLARIN-MULTISENS	Multimodal Corpora
K-OAr	Oral History Corpora
LLMs4SSH	Language Models
DR-LIB	Corpus Query Tools
CLARIN ELEXIS	Dictionaries
ACE	Sign Language Resources

Table 1: K-centres as curators of particular CRF overviews

3 Involvement

3.1 Using the GitHub-based workflow

In addition to having a more robust technical backend in general, the new setup for the CRF also provides a much easier way for collaboration. With each item of each family being contained in its own JSON file in the GitHub repository, collaborators can simply use the normal git-based way of collaboration. They can fork the repository, edit the specific items that they want to contribute to, or suggest adding new ones or deleting deprecated ones. Afterwards, they can make a pull request, which notifies the maintainers of the CRF working at the central hub, who in turn review the suggested changes. The accepted changes are immediately picked up by the CI pipeline and are reflected on the CLARIN website as well. As described in Section 2.4, the repository also stores open issues in GitHub tickets, so less tech-savvy users can simply create tickets with their suggested changes.

3.2 Domain expertise via CLARIN K-centres

Having the possibility for easy collaboration also means that it makes sense now to involve domain experts in helping with the curation of the CRF. In collaboration with the Knowledge Infrastructure Committee,²² we have started reaching out to various CLARIN Knowledge Centres that seem to have an obvious connection to one of the families. As of 29 August 2025, nine K-centres have agreed to become maintainers, each of their own resource family – they are presented in Table 1.²³

Such collaboration is beginning to ensure that the work of maintaining the CRF is shared over many people, which will help make the effort more sustainable in the long term.

4 CLARIN Resource Families in comparison to the VLO

4.1 Conceptual and technological differences

We lastly address how the CRF overviews crucially differ from but ultimately complement the VLO (Goosen & Eckart, 2014; Van Uytvanck et al., 2010, 2012), which is CLARIN ERIC’s flagship LRT discovery service.

The fundamental difference between the CRF and the VLO boils down to a conceptual as well as technological one. The listings of LRTs in the VLO originate from the periodic and crucially automatic

¹⁹<https://www.zotero.org/groups/562080/clarin/collections/GDJJGIBW/collection>

²⁰Thus far approximately 70% of the tickets have been closed, with many existing metadata records having also been updated directly because of this. This curation effort has also led to the systematic overhaul of deprecated metadata records of corpora belonging to the legacy LRT collection. An example of such metadata correction pertains to the *Croatian National Corpus* (Tadić, 2014), which initially had a non-working hyperlink to an access page where it had once been browsable.

²¹<https://github.com/clarin-eric/resource-families-issues/issues>

²²<https://www.clarin.eu/governance/knowledge-infrastructure-committee>

²³See <https://www.clarin.eu/k-centre-catalogue> for the CLARIN K-centre catalogue.

harvesting of metadata records from resource providers (Van Uytvanck et al., 2012, pp. 1029–1031). These resource providers are usually but not exclusively the repositories of CLARIN B- and C- centres. By contrast, the CRF overviews employ no such automated harvesting – the listings of LRTs in the CRF overviews are wholly manually created and then also curated by hand. While there are fewer than 2000 LRTs listed in the CRF overviews, there are almost 930,000 records in the VLO as of 1 April 2026. Further, the majority of the fewer-than-2000 LRTs listed in the CRF also constitutes a subset of the approximately 930,000 VLO records, which means that what can be found in the CRF overviews can in almost all cases be found in the VLO as well (see also Section 2.3 about how this overlap dovetails with the inclusion criteria).

The CRF overviews might thus seem redundant in view of VLO’s considerably larger offer. The existence of the CRF overviews, however, may be justified precisely on the grounds of manual curation, which remains important to this day for working with heterogenous collections of big data (Yoon et al., 2025). In the case of CLARIN, manual curation is valuable in relation to the aforementioned taxonomic division into categories such as ‘parliamentary corpora,’ ‘corpora of academic texts,’ and so on, which are not yet included in the VLO (but see Section 4.2). What is important is that such taxonomic categories are defined and demarcated on grounds that are inherently fuzzy and informal, and thus require a large degree of qualitative consideration. For instance, the Spoken corpora family does not correspond to just all corpora of audio recordings,²⁴ but on the CRF definition refers to those corpora that are primarily used by researchers for the (not necessarily computational) study of the linguistic or discursive features of spoken language production (Adolphs & Knight, 2010). What is excluded in this way is the fairly large set of smaller datasets of speech recordings that are primarily intended to be training models for developing speech technologies; such exclusion is in line with the fact that the CRF initiative is primarily aimed at researchers working in the (digital) humanities and social sciences rather than computer scientists and resource developers.

4.2 Including CRF taxonomies in the VLO

The VLO is strictly agnostic in terms of the metadata values that it harvests. The non-existence of the CRF categories in the VLO has thus far reflected the fact that the service providers (i.e., the B- and C- centres) do not specify that a given corpus or tool belongs to this or that taxonomic category in a principled fashion; for instance, as a metadatum that would end up as a dedicated CMDI component when harvested by the VLO (Windhouwer & Goosen, 2022). They likely do not do so because of the aforementioned fuzziness of the categories, and the related lack of consensus on what constitutes a particular family.

In spite of this, such taxonomic divisions remain important for directing researchers to quality datasets. Because the VLO also comes with the goal of high recall (anything relevant within the CLARIN ecosystem should be findable through it), it is the intention to let the VLO index every item that is featured in the CRF overviews. This might come at the cost of duplication, as certain resources will be shown twice in the VLO, once based on the original repository’s metadata and once based on on the CRF entry; however, this is deemed better than not being able to find entries (or at least not easily) that are included in a CRF overview. Additionally, the VLO will be able to group such duplicated entries together and therefore make it clear to the user that these are different representations of the same dataset.

4.3 CRF supporting VLO findability

The manual CRF listings contribute quite considerably to visibility. In a 2018 survey of VLO findability, Fišer and Lenardič (2018) report that by using simple queries such as *parliament* corpus*, only about 75% of the corpora that were at the time included as part of the *Parliamentary corpora* CRF could straightforwardly be found in the VLO. The difficult findability of the rest generally had to do with the fact that the relevant corpora were ranked rather low in the VLO results, sometimes even lower than non-parliamentary corpora, for reasons related to the technical ranking algorithm employed by the VLO. An advantage of the CRF overviews here is that they focus on a consistent level of LRT granularity (e.g., CRF overviews list just one version of a single corpus rather than all the versions or possible subsets, in

²⁴<https://www.clarin.eu/resource-families/spoken-corpora>

contrast to the VLO), and from 2025 onwards, only list available LRTs (see Section 2.3).

Such difficulties in VLO recall persist to this day. For instance, with the same search string as above – that is, *parliament* corpus* –, the Czech corpus *ParCzech 4.0* (Kopp, 2024) ends up only on only page 10 of the search results, and crucially gets outranked by corpora such as the Slovenian *metaFida 1.0* (Erjavec, 2023) corpus and the *Balanced Corpus of Modern Latvian (LVK2013)* (Levāne-Petrova et al., 2013). The latter are not parliamentary corpora proper but rather reference/balanced corpora that contain parliamentary proceedings as a relatively small subset. The CRF initiative thus boosts the findability of a corpus like *ParCzech 4.0* (Kopp, 2024) by including it in a rather small but visible online overview that consists of parliamentary corpora exclusively.

5 Conclusion

We have first presented the new backend of the CLARIN Resource Families (CRF) initiative, which corresponds to a Python-based workflow that takes JSON files as inputs. This backend has been developed to help open up the CRF initiative to external collaborators. We have then at some length discussed the qualitative guidelines for describing the language resources and tools constituting the CRF overviews, while highlighting certain non-trivial issues that pertain – also in a general sense – to the documentation of descriptive metadata like language and annotation. We have then moved on to discussing how potential collaborators can become involved in the initiative and the ongoing process of facilitating curatorial help via CLARIN K-centres, which have taken over management of some of the families. Finally, we have discussed how CRF relates to CLARIN ERIC’s flagship resource discovery service – that is, the Virtual Language Observatory –, from both the technological and conceptual perspectives.

Work in the near future will mostly focus on streamlining the collaboration workflow, especially in relation to the K-centres. For K-centre representatives, possible training sessions for using the workflow will be organised as needed. If new overviews are to be added to the CRF (e.g., medical corpora), a gap analysis will first be performed in collaboration with the K-centres to identify options for potential expansions.

Acknowledgments

The work in this paper has been funded by the CLARIN Resource Families project for all authors, as well as by the Slovenian Research Agency programme *P6-0436: Digital Humanities: resources, tools and methods* (2022–2027) for authors 1 and 3.

References

- Adolphs, S., & Knight, D. (2010). Building a spoken corpus: What are the basics? In A. O’Keeffe & M. J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 38–52). London: Routledge. <https://doi.org/10.4324/9780367076399-3>
- Ahačič, K., Legan Ravnkar, A., Merše, M., Narat, J., & Novak, F. (2011). *Words of the 16th-century slovenian literary language*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1127>
- Bański, P., & Hedeland, H. (2022). Standards in CLARIN. In D. Fišer & A. Witt (Eds.), *CLARIN. The infrastructure for language resources* (pp. 307–340). De Gruyter. <https://doi.org/10.1515/9783110767377-012>
- Bański, P., Heid, U., & Herzberg, L. (Eds.). (2025). *Harmonizing language data: Standards for linguistic resources*. De Gruyter. <https://doi.org/10.1515/9783112208212>
- Davies, M. (2017). *Corpus of Historical American English – Kielipankki Korp version 2017H1*. Kielipankki. <http://urn.fi/urn:nbn:fi:lb-2017061925>
- De Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational linguistics*, 47(2), 255–308. https://doi.org/10.1162/coli_a_00402
- Erjavec, T. (2012). Multext-east: Morphosyntactic resources for central and eastern european languages. *Language resources and evaluation*, 46(1), 131–142. <https://doi.org/10.1007/s10579-011-9174-8>

- Erjavec, T. (2017). Multext-east. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 441–462). Springer. https://doi.org/10.1007/978-94-024-0881-2_17
- Erjavec, T. (2023). *Corpus of combined Slovenian corpora metaFida 1.0*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1775>
- Finatto, M., & Grilo, S. (2020). *CorPop: A corpus of popular Brazilian Portuguese*. PORTULAN. <http://hdl.handle.net/21.11129/0000-000D-FE56-5>
- Fišer, D., & Lenardič, J. (2018). Parliamentary corpora in the CLARIN infrastructure. In M. Piasecki (Ed.), *Selected papers from the CLARIN annual conference 2017* (pp. 75–85). <https://doi.org/10.3384/ecp147>
- Fišer, D., Lenardič, J., & Erjavec, T. (2018). CLARIN's key resource families. *Eleventh international conference on Language Resources and Evaluation [LREC 2018]*. <https://aclanthology.org/L18-1210.pdf>
- Forkel, R., & Hammarström, H. (2022). Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web*, 13(6), 917–924. <https://doi.org/10.3233/SW-212843>
- Goosen, T., & Eckart, T. (2014). Virtual language observatory 3.0: What's new. *CLARIN annual conference 2014*. Retrieved January 12, 2026, from http://www.clarin.eu/sites/default/files/cac2014_submission_2.0.pdf
- Hajič, J., Hlaváčová, J., Mikulová, M., Straka, M., & Štěpánková, B. (2024). *MorfFlex CZ 2.1 (2024-12-23)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal; Applied Linguistics (ÚFAL). <http://hdl.handle.net/11234/1-5833>
- Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2026). *Glottolog/glottolog: Glottolog database 5.3* (Version v5.3). Zenodo. <https://doi.org/10.5281/zenodo.18840935>
- Helgadóttir, S., & Barkarson, S. (2018). *The saga corpus (Fornritin)*. CLARIN-IS. <http://hdl.handle.net/20.500.12537/32>
- Kopp, M. (2024). *ParCzech 4.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal; Applied Linguistics (ÚFAL). <http://hdl.handle.net/11234/1-5360>
- Krek, S., Erjavec, T., Repar, A., Čibej, J., Arhar Holdt, Š., Gantar, P., Kosem, I., Robnik-Šikonja, M., Ljubešić, N., Dobrovoljc, K., Laskowski, C., Grčar, M., Holozan, P., Šuster, S., Gorjanc, V., Stabej, M., & Logar, N. (2019). *Corpus of written standard Slovene Gigafida 2.0*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1320>
- Lenardič, J. (2023). Documenting corpus annotation in CMDI: State of affairs. In K. Lindén, J. Niemi, & T. Kontino (Eds.), *CLARIN annual conference proceedings* (pp. 48–52). Retrieved January 23, 2026, from <https://office.clarin.eu/v/CE-2023-2328-CLARIN2023-ConferenceProceedings.pdf#page=56>
- Lenardič, J., & Fišer, D. (2022a). CLARIN depositing guidelines: State of affairs and proposals for improvement. *CLARIN Annual Conference Proceedings*, 48–52. <https://office.clarin.eu/v/CE-2022-2118-CLARIN2022-ConferenceProceedings.pdf#page=54>
- Lenardič, J., & Fišer, D. (2022b). The CLARIN resource and tool families. In D. Fišer & A. Witt (Eds.), *CLARIN. The infrastructure for language resources* (pp. 343–372). De Gruyter. <https://doi.org/10.1515/9783110767377-013>
- Lenardič, J., & König, A. (2025). *Guidelines for preparing CLARIN resource families*. CLARIN Office Archive. <https://doi.org/10.34733/doc-187>
- Levāne-Petrova, K., Pokratiece, K. 1., Vēvere, D., Poikāns, I., & Andronova, E. (2013). *Balanced corpus of modern Latvian (LVK2013)*. CLARIN-LV digital library at IMCS, University of Latvia. <http://hdl.handle.net/20.500.12574/44>
- Ljubešić, N., & Erjavec, T. (2025). Part-of-speech tagging and related annotation. In P. Bański, U. Heid, & L. Herzberg (Eds.), *Harmonizing language data: Standards for linguistic resources* (pp. 61–88). De Gruyter. <https://doi.org/10.1515/9783112208212-004>
- Mamede, N., Baptista, J., & Mamede, N. (2020). *ViPER verb lexical database*. PORTULAN. <http://hdl.handle.net/21.11129/0000-000D-F91E-A>

- McGillivray, B. (2025). *LatinISE corpus (version 5)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal; Applied Linguistics (ÚFAL). <http://hdl.handle.net/11372/LRT-5870>
- Meden, K., Erjavec, T., & Pančur, A. (2025). Slovenian parliamentary corpus siParl. *Language Resources and Evaluation*, 59(2), 891–911. <https://doi.org/10.1007/s10579-024-09746-8>
- Nordhoff, S. (2012). Linked data for linguistic diversity research: Glottolog/Lnangdoc and ASJP online. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), *Linked data in linguistics: Representing and connecting language data and language metadata* (pp. 191–200). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28249-2_18
- Odijk, J. (2019). Discovering software resources in CLARIN. *Selected Papers of the CLARIN Conference 2018*, 159, 121–132. <http://hdl.handle.net/1874/391802>
- Pančur, A., Meden, K., Erjavec, T., Ojsteršek, M., Šorn, M., & Blaj Hribar, N. (2024). *Slovenian parliamentary corpus (1990–2022) siParl 4.0*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1936>
- Romary, L. (2025). International standards for the identification and the description of languages and their varieties. In P. Bański, U. Heid, & L. Herzberg (Eds.), *Harmonizing language data: Standards for linguistic resources* (pp. 35–60). De Gruyter. <https://doi.org/10.1515/9783112208212-003>
- Straka, M., & Straková, J. (2014). *NameTag*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal; Applied Linguistics (ÚFAL). <http://hdl.handle.net/11858/00-097C-0000-0023-43CE-E>
- Straková, J., & Straka, M. (2025, July). NameTag 3: A tool and a service for multilingual/multitagset NER. In P. Mishra, S. Muresan, & T. Yu (Eds.), *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 3: System demonstrations)* (pp. 31–39). Association for Computational Linguistics. <https://aclanthology.org/2025.acl-demo.4/>
- Tadić, M. (2014). *Croatian national corpus*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal; Applied Linguistics (ÚFAL), Faculty of Mathematics; Physics, Charles University. <http://hdl.handle.net/11372/LRT-233>
- Van Uytvanck, D., Stehouwer, H., & Lampen, L. (2012). Semantic metadata mapping in practice: The virtual language observatory. *Eight international conference on Language Resources and Evaluation [LREC 2012]*, 1029–1034. <https://doi.org/11858/00-001M-0000-000F-85EE-8>
- Van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P., & Gardelleni, M. (2010). Virtual language observatory: The portal to the language resources and technology universe. *Seventh conference on International Language Resources and Evaluation [LREC 2010]*, 900–903. <https://doi.org/11858/00-001M-0000-0012-B734-9>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3(1), 1–9. <https://doi.org/10.1038/sdata.2016.18>
- Windhouwer, M., & Goosen, T. (2022). Component metadata infrastructure. In D. Fišer & A. Witt (Eds.), *CLARIN. The infrastructure for language resources* (pp. 191–222). De Gruyter. <https://doi.org/10.1515/9783110767377-008>
- Yoon, A., Kim, J., & Donaldson, D. R. (2025). Big data curation framework: Curation actions and challenges. *Journal of Information Science*, 51(1), 205–223. <https://doi.org/10.1177/01655515221133528>
- Žak, P., & Skuczyńska, B. (2018). *OptaHopper: Phrase-level sentiment with opinion targets*. CLARIN-PL digital repository. <http://hdl.handle.net/11321/584>