Towards a new CLARIN Resource Family: Tools for Manual Linguistic Annotation

Kaja Dobrovoljc Zor
Jožef Stefan Institute
University of Ljubljana
Ljubljana, Slovenia
kaja.dobrovoljc@ijs.si

Ludovica Pannitto
University of Bologna
Bologna, Italy
ludovica.pannitto@unibo.it

Bruno Guillaume
Université de Lorraine, CNRS, Inria, LORIA
Nancy, France
bruno.guillaume@loria.fr

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

Abstract

Manual annotation tools are essential infrastructure for the development and enhancement of language resources, yet have so far been overlooked in the CLARIN Resource Families (CRF) initiative, which covers tools primarily for automatic processing. To bridge this gap, we present a new CRF dedicated to tools for manual linguistic annotation, describe the process of identifying relevant tools and defining their descriptions within the CRF metadata schema, discuss limitations encountered along the way, and outline directions for future expansion.

1 Introduction

Manual annotation tools are core infrastructure for language resource development, enabling linguistically informed corpora that underpin linguistic research (Gries & Berez, 2017) and NLP applications (Ide & Pustejovsky, 2017; Wilcock, 2009), including the training, evaluation and probing of large language models (Felkner et al., 2024; Liu et al., 2023; Stollenwerk et al., 2023). Despite their central role, such tools have received limited systematic attention within the CLARIN infrastructure: while the CLARIN Resource Families (CRF) initiative (Fišer et al., 2018; Lenardič & Fišer, 2022) provides curated, community-maintained overviews of language resources and tools available across CLARIN repositories, existing tool-focused families address primarily automatic processing tasks such as named entity recognition, normalisation, and part-of-speech tagging. Tools supporting manual annotation workflows have so far not been represented.

To address this gap, we present a new CLARIN Resource Family dedicated to tools for manual linguistic annotation. The family is derived from a recent comprehensive survey of 21 tools, primarily aimed at morphosyntactic and multi-word expression annotation (Pannitto et al., 2026), conducted within the UniDive COST Action (Savary et al., 2024), which systematically documented over 50 features relevant to manual annotation workflows, and made its results available as an open dataset and interactive online platform. From this survey, we identified the subset of tools affiliated with the CLARIN infrastructure and complemented this with a systematic search of additional relevant tools, resulting in a curated family of **11 tools** openly accessible at https://doi.org/10.34733/crf-manual-annotation (development version; public release forthcoming).

The remainder of this paper describes the methodological decisions involved in constructing the family (Section 2), discusses the outcomes and challenges encountered (Section 3), and outlines directions for future work.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

2 From Survey to Resource Family

2.1 Tool Selection

Pannitto et al. (2026) recently conducted a comprehensive survey of tools for manual linguistic annotation, motivated by the lack of systematic, fine-grained documentation of annotation tool features relevant to researchers working with diverse languages and annotation tasks. The survey was implemented as a structured questionnaire distributed both directly to tool development teams and broadly within the UniDive COST Action community (Savary et al., 2024), and documented over 50 features across 11 thematic categories — from software infrastructure and language support to workflow and annotation analysis. Although primarily designed to assess support for morphosyntactic annotation (e.g. lemmatisation, part-of-speech tagging, and dependency parsing as in Universal Dependencies) and multi-word expression annotation (e.g. as in PARSEME), the resulting set of 21 tools mostly supports annotation of a much broader range of phenomena, from named entity recognition and coreference to semantic roles and discourse structure. The survey results were published as an open dataset and interactive online platform.[1]

As the most recent and in-depth survey of its kind, this work represented the most sensible starting point for constructing the new CRF, which we complemented with a systematic search of existing online tool inventories and surveys — including MALA (Gessler, 2022) and the corpus-analysis.com catalogue (Berberich & Kleiber, 2023) — and the Virtual Language Observatory (VLO), the CLARIN infrastructure's main resource discovery portal. Throughout this process, tool eligibility was assessed against three criteria: (1) **manual annotation support**, i.e. the tool must provide an interactive environment for human annotators rather than performing automatic processing; (2) **general-purpose annotation support**, i.e. the tool must support multiple linguistic annotation types or configurable annotation schemes, rather than being designed exclusively for a single specialised task (e.g. coreference annotation, speech segmentation, or OCR correction only); and (3) **CLARIN affiliation**: as the CRF initiative is designed to promote tools and resources within the CLARIN ecosystem rather than serve as a general-purpose inventory (Lenardič et al., 2026, Section 2.3), each tool must have a demonstrable connection to the CLARIN infrastructure. While for corpora and lexical resources this normally requires a deposit or metadata record in a CLARIN B- or C-centre, the CRF guidelines (Lenardič & König, 2025, p. 7) allow greater flexibility for software, which is less systematically represented in CLARIN repositories (Odijk, 2019): tools without such a record may therefore be included where another explicit CLARIN connection can be established (e.g., through the author(s)' affiliation).[2]

This process resulted in 11 tools to be included in the newly introduced CRF of tools for manual linguistic annotation: **ELAN** (Max Planck Institute, Wittenburg et al. 2006), **EXMARaLDA** (HZSK, Schmidt 2012), **FLAT** (CLARIAH-NL, van Gompel and Reynaert 2013), **FLeX** (LINDAT, Butler and Van Volkinburg 2007), **INCEpTION** (CLARIN:EL, Eckart De Castilho et al. 2024), **Q-CAT** (CLARIN.SI, Brank 2023), **TrEd** (LINDAT, Pajas and Fabian 2019), **WebAnno** (CLARIN-D, Eckart de Castilho et al. 2016), **ANNIS** (CLARIN-D, Krause and Zeldes 2015), **TEITOK** (LINDAT, Janssen 2021), and **TreeTagger** (CLARIN-D, Schmid 1994).

2.2 Tool Description

For each tool, we defined a concise set of features to include in the CRF entries, selecting those most relevant to end-users and sufficiently discriminating across tools; features with near-uniform values were omitted as uninformative (e.g. tagset customisation, supported by large majority of the tools). The resulting schema includes: tool **name** and **URL**; a brief **description** of the tool as advertised online; **licence** (e.g. Apache 2.0); **software type** (desktop- or web-based); **input formats** (e.g. raw text, CoNLL-U, XML); **annotation types** (e.g. tokens, spans, relations, chains); and key **publication**(s). For the large

[1] https://unidive.github.io/corpus-annotation-tools/interface/tools.html

[2] Of the 21 tools included in the UniDive survey, eight are represented in the current CRF; the remaining 13 — Arborator-Grew, BoAT, brat, CoNLLU-Editor, DgAnnotator, GitDox, Label Sleuth, Label Studio, Palmyra, PLURAL, TALEN, Toolbox, and UD Annotatrix — were not selected under the eligibility criteria applied at the time of compilation. Their inclusion may be reconsidered in future revisions of the family.

majority of tools originating from the UniDive survey, these values were obtained automatically by conversion from the published TSV dataset; for the 3 tools identified through additional searches, entries were compiled manually following the same schema.

Users requiring more fine-grained information are referred to the external resources listed in the CRF family introduction, including the UniDive interactive platform (See Footnote 1) specifically designed for that purpose.

2.3 Online Publication

The resulting entries were encoded as JSON files following the CRF guidelines (Lenardič & König, 2025) and submitted to the CLARIN Resource Families GitHub repository.[3] The family is publicly accessible at https://doi.org/10.34733/crf-manual-annotation, with the backend code available at https://github.com/clarin-eric/clarin-resource-families/tree/main/tools/manual-annotation-tools, where they can be freely edited and extended via pull requests.

3 Discussion

The new CRF for tools for manual linguistic annotation fills a gap in the CLARIN infrastructure, enabling straightforward discovery of relevant tools for researchers in digital humanities, social sciences, and language technology. Its construction, however, also surfaced a number of challenges worth noting.

First, establishing the required CLARIN connection (Lenardič et al., 2026, Section 2.3) proved non-trivial: very few relevant results — for example, only 7 of the 11 tools included in the family can also be found through the VLO — reflecting the fact that software is rarely deposited systematically in CLARIN repositories and more commonly hosted on platforms such as GitHub. Second, where tools were deposited, metadata quality was often insufficient to determine the nature and scope of annotation supported — for instance, whether annotation is manual or automatic, or whether the tool is general-purpose or restricted to a specific annotation task. Third, several candidates had to be excluded due to poor or outdated metadata that obscured critical information about the tool's current status, such as discontinued maintenance, broken links, or insufficient installation documentation. An example of an excluded tool is *Atomic*, in which case even the handle link listed in the VLO returns an '404 error' on 3 April 2026; see Druskat et al. (2015).

These observations point to a broader need for improved software metadata standards within the CLARIN infrastructure — an issue already raised by Borisova et al. (2024), who found that annotation tools are generally accessible and interoperable but difficult to find and reuse, with most lacking persistent identifiers, semantically structured metadata, and detailed provenance information. Our experience confirms that these shortcomings directly hinder integration into curated infrastructure overviews such as the CRF, and we therefore recommend that tool authors deposit their software in a CLARIN B- or C-centre to ensure discoverability via the VLO, and provide structured metadata explicitly describing the type and scope of annotation supported, including version and citation information, with deposited links and installation instructions kept up to date. Another issue is that it is often quite unclear how to cite the tools that do not have corresponding metadata records in a CLARIN B- or C- centre in a manner that is compliant with the CLARIN citation guidelines (Matthiesen & Lenardič, 2025). For instance, the only instruction for citing TEITOK is to cite an old paper describing the tool (i.e., Janssen, 2016), but there are no citation instructions on how to cite the tool itself.[4]

4 Conclusion

We have presented a new CLARIN Resource Family for tools for manual linguistic annotation — the first CRF dedicated to manual data annotation — based on a systematic selection process involving state-of-the-art surveys and targeted infrastructure searches. The family currently comprises 11 tools and is designed to grow through community contributions via the GitHub-based CRF workflow. It is

[3] https://github.com/clarin-eric/clarin-resource-families/tree/main

[4] See http://www.teitok.org/index.php?action=credits.

conceived as a first step towards broader coverage of manual corpus annotation tools within the CLARIN ecosystem, and we invite the community to help expand and maintain it.

Acknowledgments

This work was supported by the COST Action CA21167 UniDive (European Cooperation in Science and Technology), EU ERA Chair grant no.101186647 (AI4DH), and ARIS grants P6-0411, P6-0436 and GC-002. We also acknowledge support from CLARIN.SI and the CLARIN Resource Families initiative.

References

Berberich, K., & Kleiber, I. (2023). Tools for corpus linguistics [Accessed: 2025-10-15].

Borisova, E., Abu Ahmad, R., Garcia-Castro, L., Usbeck, R., & Rehm, G. (2024, March). Surveying the FAIRness of annotation tools: Difficult to find, difficult to reuse. In S. Henning & M. Stede (Eds.), *Proceedings of the 18th linguistic annotation workshop (law-xviii)* (pp. 29–45). Association for Computational Linguistics. https://aclanthology.org/2024.law-1.4/

Brank, J. (2023). Q-CAT corpus annotation tool 1.5 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1844

Butler, L., & Van Volkinburg, H. (2007). FieldWorks Language Explorer (FLEx). *Technology Review*, *1*(1), 1.

Druskat, S., Bierkandt, L., Gast, V., Rzymski, C., & Zipser, F. (2015). *Atomic - multilevel corpus annotation tool*. IMS, CLARIN-D Centre, University of Stuttgart. http://hdl.handle.net/11022/1007-0000-0000-8E9D-0

Eckart De Castilho, R., Klie, J.-C., & Gurevych, I. (2024, November). Integrating INCEpTION into larger annotation processes. In D. I. Hernandez Farias, T. Hope, & M. Li (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing: System demonstrations* (pp. 110–121). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-demo.12

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 76–84. https://www.aclweb.org/anthology/W16-4011

Felkner, V., Thompson, J., & May, J. (2024, August). GPT is not an annotator: The necessity of human annotation in fairness benchmark construction. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 14104–14115). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.760

Fišer, D., Lenardič, J., & Erjavec, T. (2018). Clarin's key resource families. *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Gessler, L. (2022). MALA: The map of applications for linguistic annotation [Software available from https://github.com/lgessler/mala]. https://github.com/lgessler/mala

Gries, S. T., & Berez, A. L. (2017). Linguistic annotation in/for corpus linguistics. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 379–409). Springer. https://doi.org/10.1007/978-94-024-0881-2_15

Ide, N., & Pustejovsky, J. (2017). *Handbook of linguistic annotation* (1st). Springer.

Janssen, M. (2016). Teitok: Text-faithful annotated corpora. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4037–4043.

Janssen, M. (2021). Integrating TEITOK and KonText/PMLTQ at LINDAT. *Selected Papers from the CLARIN Annual Conference 2020*, 104–110. https://doi.org/10.3384/ecp18012

Krause, T., & Zeldes, A. (2015). *Annis search and visualization system for corpora*. IMS, CLARIN-D Centre, University of Stuttgart. http://hdl.handle.net/11022/1007-0000-0000-8E9C-1

Lenardič, J., & Fišer, D. (2022). The clarin resource and tool families. In D. Fišer & A. Witt (Eds.), *The infrastructure for language resources* (pp. 343–372). De Gruyter. https://doi.org/doi:10.1515/9783110767377-013

Lenardič, J., & König, A. (2025). *Guidelines for preparing CLARIN resource families* (tech. rep. No. CE-2024-2451). CLARIN ERIC. https://office.clarin.eu/v/CE-2024-2451-guidelines-for-preparing-clarin-resource-families.pdf

Lenardič, J., König, A., Pahor de Maiti Tekavčič, K., & Van Uytvanck, D. (2026). The CLARIN resource families workflow. In C. Grisot & T. Kontino (Eds.), *Selected papers from the CLARIN annual conference 2025* (pp. 89–99). https://doi.org/10.3384/ecp222.1602

Liu, Y., Fabbri, A., Liu, P., Zhao, Y., Nan, L., Han, R., Han, S., Joty, S., Wu, C.-S., Xiong, C., & Radev, D. (2023, July). Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 4140–4170). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.228

Matthiesen, M., & Lenardič, J. (2025). CLARIN data citation guidelines. https://doi.org/10.34733/DOC-189

Odijk, J. (2019). Discovering software resources in CLARIN. *Selected Papers of the CLARIN Conference 2018*, *159*, 121–132. http://hdl.handle.net/1874/391802

Pajas, P., & Fabian, P. (2019). Tree Editor TrEd [Software tool. Last released version 2.5236 (Feb 28, 2019).].

Pannitto, L., Dobrovoljc, K., & Guillaume, B. (2026). Survey of tools for manual linguistic annotation: Supporting diversity through interactive exploration. *Proceedings of the Fifteenth Language Resources and Evaluation Conference*.

Savary, A., Zeman, D., Barbu Mititelu, V., Barreiro, A., Caftanatov, O., de Marneffe, M.-C., Dobrovoljc, K., Eryiğit, G., Giouli, V., Guillaume, B., Markantonatou, S., Melnik, N., Nivre, J., Ojha, A. K., Ramisch, C., Walsh, A., Wójtowicz, B., & Wróblewska, A. (2024, May). UniDive: A COST action on universality, diversity and idiosyncrasy in language technology. In M. Melero, S. Sakti, & C. Soria (Eds.), *Proceedings of the 3rd annual meeting of the special interest group on under-resourced languages @ lrec-coling 2024* (pp. 372–382). ELRA; ICCL. https://aclanthology.org/2024.sigul-1.45/

Schmid, H. (1994). *TreeTagger*. IMS, CLARIN-D Centre, University of Stuttgart. http://hdl.handle.net/11022/1007-0000-0000-8E4D-B

Schmidt, T. (2012, May). EXMARaLDA and the FOLK tools — two toolsets for transcribing and annotating spoken language. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 236–240). European Language Resources Association (ELRA). https://aclanthology.org/L12-1295/

Stollenwerk, F., Öhman, J., Petrelli, D., Wallerö, E., Olsson, F., Bengtsson, C., Horndahl, A., & Gandler, G. Z. (2023). Text annotation handbook: A practical guide for machine learning projects. https://arxiv.org/abs/2310.11780

van Gompel, M., & Reynaert, M. (2013). Folia: A practical xml format for linguistic annotation – a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, *3*, 63–81. https://clinjournal.org/clinj/article/view/26

Wilcock, G. (2009). *Introduction to linguistic annotation and text analytics*. Springer International Publishing.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006, May). ELAN: A professional framework for multimodality research. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the fifth international conference on language resources and evaluation (LREC'06)*. European Language Resources Association (ELRA). https://aclanthology.org/L06-1082/