Open Research Award 2021 nomination

Nominees: Andrea Abel, Lavinia Nicoleta Aparaschivei, Greta Franzini, Jennifer-Carmen Frey, Verena Lyding, Lionel Nicolas, Egon W. Stemle (*Language Technologies* research group).

The growing *Language Technologies* (LT) group at the Institute of Applied Linguistics actively pursues Computational Linguistics and Digital Humanities research related to the creation, curation, adaptation, and long-term preservation of language resources and tools, and provides computational expertise and technical support to the research efforts of the entire Institute. In so doing, LT enhances traditional research design and procedures with the latest developments from Computational Linguistics and Natural Language Processing, while better understanding and accommodating valuable information on the practices and needs of linguistic research stakeholders.

LT's purview stretches across disciplines (e.g., humanities, social sciences, computer science), languages (e.g., Italian, German, English, Ladin, Cimbrian, Czech, French, Dutch), geographical boundaries and communities (e.g., researchers, language-related professionals, pupils, citizens), and manifests itself in the active participation and/or coordination of initiatives designed to bring people together (e.g., *enetCollect*), invite them to join in the research (e.g., *Zeit.shift*) and help shape best practice (e.g., (meta)data standardization for learner corpora).

To reach scientific and lay audiences alike, and improve openness, transparency, inclusion and reproducibility within and beyond the walls of the Institute of Applied Linguistics, LT is and always has been committed to open science. It fully subscribes to the FAIR guiding principles for scientific data management and stewardship, and fosters their application through training workshops and publications on related efforts and best practices [4, 5, 6, 8, 9][1]. More specifically, LT employs and customizes open source software and platforms, releases tools, data and code --along with README files, changelogs and versioning-- under open licenses and endeavors to publish scientific articles in open access. Additionally, LT hosts and maintains European and local digital infrastructures (e.g., *CLARIN*, *DI-ÖSS*, *PORTA*) to increase the visibility and promote the use of the Institute's research output.

Open datasets prepared and released by LT include the *MERLIN* written learner corpus for Czech, German, and Italian [2][2]; the *DiDi* multilingual corpus of Facebook posts, comments and private messages written by 136 South Tyrolean users, unique for its sociolinguistic metadata [3][3]; the *KoKo* error-annotated learner corpus of L1 German speakers [1][4]; the *Kolipsi* corpus family of South Tyrolean L2 Italian and German learners [19][5]; the *LEONIDE* longitudinal corpus of student essays documenting the language

[1] https://www.go-fair.org/fair-principles/
[2] https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/6
[3] https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/7
[4] https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/12
[5] https://www.porta.eurac.edu/lci/kolipsi-family/

competences and writing development of lower secondary school students in South Tyrol [7][6]; and the *PAISÀ* corpus of contemporary Italian [13][7], which has become a standard resource for Italian language tasks.

Among the many open source software packages, tools and applications developed by the group, noteworthy are the Italian extension of the *Common Text Analysis Platform*[8] for the analysis of linguistic complexity [17][9]; a customized version of the *ANNIS* search and visualization architecture for use with the Institute's annotated linguistic corpora[10]; the *Transc&Anno* tool, adapted from the *FromThePage* software for the transcription and annotation of learner corpora [18][11]; the *PAISÀ* interface hosting the previously mentioned *PAISÀ* corpus[12]; a keyword generator and extractor service for the online news portal *salto.bz* for short articles published in German and Italian [14][13]; a Python module to interact and retrieve data from the *DiDi* corpus through the command line[14]; and dockerized services as the 'next generation' isolation technology to encapsulate tools and applications in a reproducible manner.[15] Parallel to its own development work, LT also contributes to third-party-maintained software, namely *Lexonomy*, which is an open source, cloud-based system for writing and publishing dictionaries[16].

LT's open data infrastructures include *PORTA*, a portal to learner corpora and learner corpus infrastructure [4][17]; the *EURAC Research CLARIN Centre* --a branch of the European CLARIN language resource infrastructure[18]-- focusing on data collected at the Institute of Applied Linguistics while open to data deposits from external collaborators (e.g., from the *VinKo* project[19]) [15][*20*]; the newly-established *CLARIN Knowledge Centre for Computer-Mediated Communication and Social Media*, which provides expertise, support and training in this domain for the production, modification and publishing of relevant resources and technologies[21]; and the *Digital Infrastructure for the ecosystem of South Tyrolean language data and services (DI-ÖSS)*, which, as the name suggests, seeks to establish and optimize data and service exchange processes between South Tyrolean news and memory institutions [10, 11][22]**.**

---

[6] https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/25
[7] https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/3
[8] http://sifnos.sfs.uni-tuebingen.de/ctap/
[9] https://github.com/commul/ctap
[10] https://github.com/commul/ANNIS
[11] https://github.com/commul/transcanno
[12] http://corpusitaliano.it/
[13] https://gitlab.inf.unibz.it/commul/di-oss/api-service-salto
[14] https://gitlab.inf.unibz.it/commul/didi/
[15] https://gitlab.inf.unibz.it/commul/docker
[16] https://www.lexonomy.eu/docs/intro and https://github.com/elexis-eu/lexonomy
[17] www.porta.eurac.edu
[18] https://www.clarin.eu/
[19] https://www.vinko.it/project.php
[20] https://clarin.eurac.edu/
[21] https://cmc-corpora.org/dev/ckcmc/
[22] https://www.eurac.edu/en/institutes-centers/institute-for-applied-linguistics/projects/di-oess

LT research output is openly available via the GitHub and GitLab development platforms at https://github.com/commul and https://gitlab.inf.unibz.it/commul, respectively. LT's GitHub space hosts adaptations of existing third-party tools, while GitLab is reserved for home-grown projects.

In line with Europe's strategic research priorities, LT is increasingly oriented towards participatory methodologies as a means of accelerating research progress, widening perspectives and democratizing the creation of (scientific) knowledge. In this context, LT coordinates the *enetCollect* COST Action [12, 16][23], which brings together more than one hundred stakeholders from Europe and beyond to stimulate a large-scale and long-term global research and innovation trend on language learning and crowdsourcing [12], and cooperates on the *Zeit.shift* project[24], a new (inter)national collaboration seeking to increase online access, preserve, enrich and communicate the cultural and textual heritage of the historical region of Tyrol with the help of the local population. As a "distributed intelligence" citizen science/humanities project, *Zeit.shift* does not include input from citizens at each step of the research but relies on their cognitive and observation abilities to crowdsource research data, granting them the possibility of influencing changes in project methodology, objectives, development, results and/or dissemination.

The adoption and catalyzing effect of open research practices has widened LT's readership, promoted the (re)use and citation of its output, and attracted a diverse network of collaborators, raising its global visibility. These practices also impact LT's own data collection, creation and sharing workflows, which become more forward-thinking, rigorous and reproducible with every new project.

References

1. Abel, A., Glaznieks, A., Nicolas, L., & Stemle, Egon W. (2014). KoKo: An L1 Learner Corpus for German. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 2414–2421. http://www.lrec-conf.org/proceedings/lrec2014/pdf/934_Paper.pdf
2. Boyd, A., Hana, J., Nicolas, L., Meurers, W. D., Wisniewski, K., Abel, A., Schöne, K., Stindlová, B., & Vettori, C. (2014). The MERLIN corpus: Learner language and the CEFR. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 1281–1288. http://www.lrec-conf.org/proceedings/lrec2014/pdf/606_Paper.pdf
3. Frey, J.-C., Glaznieks, A., & Stemle, Egon W. (2016). The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts. In A. Corazza, S. Montemagni, & G. Semeraro (Eds.), *Proceedings of the Third Italian Conference on Computational Linguistics* (pp. 157–161). Accademia University Press. https://www.aaccademia.it/scheda-libro?aaref=869
4. Frey, J.-C., König, A., & Fišer, D. (2020). Creating a learner corpus infrastructure: Experiences from making learner corpora available. *Proceedings of the ICTeSSH Conference 2020*, *33*, 03006. https://doi.org/10.1051/itmconf/20203303006
5. Frey, J.-C., König, A., & Stemle, E. W. (2019). How FAIR are CMC Corpora? In J. Longhi & C. Marinica (Eds.), *Proceedings 7th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora* (pp. 26–30). Cergy-Pontoise University. https://cmccorpora19.sciencesconf.org/data/pages/proceedingsCMC_Corpora2019.pdf

[23] https://enetcollect.eurac.edu/

[24] https://all4ling.eurac.edu/zeitshift/; to be launched in September 2021.

6. Frey, J.-C., König, A., Stemle, E. W., Falaise, A., Fišer, D., & Lüngen, H. (2020). The FAIR Index of CMC Corpora. In J. Longhi & C. Marinica (Eds.), *CMC Corpora through the prism of digital humanities* (pp. 127–145). L'Harmattan. https://hal.archives-ouvertes.fr/hal-03121698
7. Glaznieks, A., Frey, J.-C., Stopfner, M., Zanasi, L., & Nicolas, L. (2022). LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research*, *8*(1).
8. Glaznieks, A., Nicolas, L., Stemle, E. W., Abel, A., & Lyding, V. (2014). Establishing a Standardised Procedure for Building Learner Corpora. *Apples - Journal of Applied Language Studies*, *8*(3), 5–20.
9. König, A., Frey, J.-C., & Stemle, E. W. (2021). Exploring Reusability and Reproducibility for a Research Infrastructure for L1 and L2 Learner Corpora. *Information*, *12*(5). https://doi.org/10.3390/info12050199
10. Lyding, V., König, A., Gorgaini, E., Nicolas, L., & Pretti, M. (2019). DI-ÖSS - Building a digital infrastructure in South Tyrol. In I. Skadina & M. Eskevich (Eds.), *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018* (Vol. 159, pp. 92–102). Linköping Electronic Conference Proceedings. https://ep.liu.se/ecp/article.asp?issue=159&article=010&volume=0
11. Lyding, V., König, A., & Pretti, M. (2020). Digital Language Infrastructures – Documenting Language Actors. *Proceedings of the 12th Language Resources and Evaluation Conference*, 3457–3462. https://aclanthology.org/2020.lrec-1.424
12. Lyding, V., Nicolas, L., Bédi, B., & Fort, K. (2018). Introducing the European NETwork for COmbining Language LEarning and Crowdsourcing Techniques (enetCollect). *Future-Proof CALL: Language Learning as Exploration and Encounters – Short Papers*, 176–181. https://doi.org/10.14705/rpnet.2018.26.833
13. Lyding, V., Stemle, E. W., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., Dittmann, H., Lenci, A., & Pirrelli, V. (2014). The PAISÀ Corpus of Italian Web Texts. *Proceedings of the 9th Web as Corpus Workshop Co-Located with the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 36–43. https://doi.org/10.3115/v1/W14-0406
14. Lyding, V., Stemle, Egon W., & König, A. (2021). Opening language resource infrastructures to non-research partners: Practicalities and challenges. In *Book of abstracts for the CLARIN Annual Conference 2021*.
15. Nicolas, L., König, A., Monachini, M., Del Gratta, R., Calamai, S., Abel, A., Enea, A., Biliotti, F., & Quochi, V. (2018). CLARIN-IT: State of Affairs, Challenges and Opportunities. *Selected Papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017*, *147*, 1–14. https://www.semanticscholar.org/paper/CLARIN-IT%3A-State-of-Affairs%2C-Challenges-and-Lionel-Alexander/e43675010666e9efbae32e4872f3d8bf43661660
16. Nicolas, L., Lyding, V., Borg, C., Forascu, C., Fort, K., Zdravkova, K., Kosem, I., Čibej, J., Arhar Holdt, Š., Millour, A., König, A., Rodosthenous, C., Sangati, F., ul Hassan, U., Katinskaia, A., Barreiro, A., Aparaschivei, L., & HaCohen-Kerner, Y. (2020). Creating Expert Knowledge by Relying on Language Learners: A Generic Approach for Mass-Producing Language Resources by Combining Implicit Crowdsourcing and Language Learning. *Proceedings of LREC 2020, 12th Language Resources and Evaluation Conference*, 268–278. https://www.aclweb.org/anthology/2020.lrec-1.34
17. Okinina, N., Frey, J.-C., & Weiss, Z. (2020). CTAP for Italian: Integrating Components for the Analysis of Italian into a Multilingual Linguistic Complexity Analysis Tool. *Proceedings of the 12th Language Resources and Evaluation Conference*, 7123–7131. https://aclanthology.org/2020.lrec-1.880
18. Okinina, N., Nicolas, L., & Lyding, V. (2018). Transc&Anno: A Graphical Tool for the Transcription and On-the-Fly Annotation of Handwritten Documents. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 701–705. https://aclanthology.org/L18-1112

19. Vettori, C., & Abel, A. (2017). *KOLIPSI II. Gli studenti altoatesini e la seconda lingua: Indagine linguistica e psicosociale. / Die Südtiroler SchülerInnen und die Zweitsprache: eine linguistische und sozialpsychologische Untersuchung.* https://doi.org/10.13140/RG.2.2.24248.96001