Progress of the PRINCIPLE Project: Promoting MT for Croatian, Icelandic, Irish and Norwegian

Petra Bago,[1] **Jane Dunne,**[2] **Federico Gaspari,**[2] **Andre Kåsen,**[3] **Gauti Kristmannsson,**[4]
Helen McHugh,[2] **Jon Arild Olsen,**[3] **Dana D. Sheridan,**[5] **Páraic Sheridan,**[5]
John Tinsley,[5] **Andy Way**[2]

[1] Faculty of Humanities and Social Sciences, University of Zagreb, 10000 Zagreb, Croatia
[2] ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland
[3] National Library of Norway, Henrik Ibsens gate 110, 0203 Oslo, Norway
[4] University of Iceland, Saemundargata 2, 101 Reykjavik, Iceland
[5] Iconic Translation Machines, Invent Building, Dublin City University, Dublin 9, Ireland

pbago@ffzg.hr, {jane.dunne,federico.gaspari,helen.mchugh,
andy.way}@adaptcentre.ie, {andre.kasen,jon.olsen}@nb.no,
gautikri@hi.is, {dana,paraic,john}@iconictranslation.com

Abstract

This paper updates the progress made on the PRINCIPLE project, a 2-year action funded by the European Commission under the Connecting Europe Facility (CEF) programme. PRINCIPLE focuses on collecting high-quality language resources for Croatian, Icelandic, Irish and Norwegian, which have been identified as low-resource languages, especially for building effective machine translation (MT) systems. We report initial achievements of the project and ongoing activities aimed at promoting the uptake of neural MT for the low-resource languages of the project.

1 Background

PRINCIPLE is a 2-year initiative that started in September 2019, funded by the European Commission under the Connecting Europe Facility (CEF) programme. The project is coordinated by the ADAPT Centre at Dublin City University (DCU, Ireland), and the consortium includes the Faculty of Humanities and Social Sciences of the University of Zagreb (Croatia), the National Library of Norway in Oslo, the University of Iceland in Rejkyavik, and Iconic Translation Machines (Ireland). PRINCIPLE focuses on the identification, collection and processing of high-quality language resources (LRs) for Croatian, Icelandic, Irish, and Norwegian (covering both varieties of Bokmål and Nynorsk), which are severely under-resourced. The uptake of machine translation (MT) for these languages has been hampered so far by the lack of extensive high-quality LRs that are required to build effective systems, especially parallel corpora. PRINCIPLE aims to improve LR collection efforts in the respective languages, prioritising the two strategic Digital Service Infrastructures (DSIs)[1] of eJustice and eProcurement. The LRs assembled and curated in PRINCIPLE will be validated to demonstrate improved MT quality, and will be uploaded via ELRC-SHARE to enhance MT systems provided by eTranslation, that are available to public administrations in Europe, thus promoting language equality for low-resource languages.

Way and Gaspari (2019) introduced the PRINCIPLE project at its start, giving a high-level overview of its main objectives, along with the planned activities and the overall approach to data collection and validation. They also explained its position within the wider eco-system of related, recently finished CEF projects such as iADAATPA (Castilho et al., 2019) and ELRI.[2] This paper provides an update on the progress of PRINCIPLE, focusing on its initial achievements and describing ongoing activities, especially in terms of engaging with stakeholders and MT users, and concludes with future plans to promote the continued collection of LRs with a view to improving and extending MT use.

[1] https://ec.europa.eu/digital-single-market/en/news/connecting-europe-facility-cef-digital-service-infrastructures
[2] www.elri-project.eu

2 Achievements and Ongoing Activities

Most of PRINCIPLE's key activities are already well underway and will continue into the second year of the action. Such activities include coordination; use-case analysis, data requirements and data preparation; development, evaluation and deployment of MT systems; identification, collection and consolidation of LRs; and, finally, dissemination. In contrast, most of the work on exploitation and sustainability will take place from July 2020, when more mature results will be available to share with the community.

The progress of the project is constantly monitored via milestones that help to verify that the work is on track to achieve the goals of the action. At the time of writing, in its first eight months, PRINCIPLE has already achieved three of these milestones, namely (i) the adoption of the jointly written project implementation plan that defines key aspects of how the action operates, (ii) the preparation of a detailed report on the use-cases, data requirements and data preparation agreed with various stakeholders in each participating country, and more recently (iii) the design of the software architecture of the MT systems to be built by Iconic for the early adopters (EAs) as part of the project, with a view to releasing the first batch by October 2020.

The EAs are data contributors who have also agreed to work with the PRINCIPLE consortium to introduce custom-built MT engines developed by Iconic into their translation workflow on the basis of their defined use-cases and requirements. These domain-adapted MT systems will be evaluated by DCU with flexible protocols including state-of-the-art automatic metrics and human/manual techniques matching the use-cases and scenarios that have been previously defined. This is a crucial step to validate the quality of the LRs in real user settings, thereby confirming their value to the actual improvement of MT system development, before uploading the data sets collected in the project to ELRC-SHARE.

As part of this ongoing effort, the consortium has already established, and continues to seek, partnerships with public and government bodies as well as private entities in Croatia, Iceland, Ireland and Norway who can provide valuable LRs to the project. By way of example, at the time of writing three data providers in Croatia have confirmed that they will be contributing LRs, thereby offering access to over 1,250 documents of varying length, and containing data from various DSIs. The Croatian PRINCIPLE partner is currently inspecting these documents to check their relevance to eJustice and eProcurement as well as their quality, before selecting those that are deemed suitable for further processing.

3 Future Plans

Preliminary indications suggest that the project is on track to achieve its ambitious objectives for LR collection, and the consortium partners will continue to collaborate with existing and new data contributors in Croatia, Iceland, Ireland and Norway to focus on the eJustice and eProcurement DSIs that are prioritised by the action.

Work to be done in the remainder of the project includes exploring the feasibility of creating new National Relay Stations[3] for the coordinated collection of LRs in Croatia, Iceland and Norway, after their successful introduction by the ELRI project. Finally, future plans also involve expanding and consolidating the LRs collected by the project, and intensifying dissemination activities, with a series of workshops scheduled in the four countries represented in PRINCIPLE by July 2021.

Acknowledgements

PRINCIPLE is generously co-financed by the European Union Connecting Europe Facility under Action 2018-EU-IA-0050 with grant agreement INEA/CEF/ICT/A2018/1761837.

References

Castilho, Sheila, Natalia Resende, Federico Gaspari, Andy Way, Tony O'Dowd, Marek Mazur, Manuel Herranz, Alex Helle, Gema Ramírez-Sánchez, Victor Sánchez-Cartagena, Mārcis Pinnis, and Valters Sics. 2019. Large-scale Machine Translation Evaluation of the iADAATPA Project. *Proceedings of Machine Translation Summit XVII, Volume 2,* Dublin, Ireland 179-185.

Way, Andy, and Federico Gaspari. 2019. PRINCIPLE: Providing Resources in Irish, Norwegian, Croatian and Icelandic for the Purposes of Language Engineering. *Proceedings of Machine Translation Summit XVII, Volume 2,* Dublin, Ireland 112-113.

[3] www.elri-project.eu/resources/D1.3_ELRI_ Public_Final_Report.pdf