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

Petra Bago,[1] **Sheila Castilho,**[2] **Jane Dunne,**[2] **Federico Gaspari,**[2] **Andre Kåsen,**[3] **Gauti Kristmannsson,**[4] **Jon Arild Olsen,**[3] **Natalia Resende,**[2] **Níels Rúnar Gíslason,**[4] **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, 102 Reykjavik, Iceland
[5] Iconic Translation Machines, Invent Building, Dublin City University, Dublin 9, Ireland

Abstract

This paper provides an overview of the main achievements of the completed PRINCIPLE project, a 2-year action funded by the European Commission under the Connecting Europe Facility programme. PRINCIPLE focused on collecting high-quality language resources for Croatian, Icelandic, Irish and Norwegian, which are severely low-resource languages, especially for building effective machine translation (MT) systems. We report the achievements of the project, primarily in terms of the large amounts of data collected for all four low-resource languages, and of promoting the uptake of neural MT for these languages.

1. Background

PRINCIPLE was a 2-year EU-funded project that ran between 2019 and 2021 to identify, collect and curate high-quality language resources (LRs) for the under-resourced languages of Croatian, Irish, Norwegian and Icelandic. The action was coordinated by the ADAPT Centre at Dublin City University (DCU), and involved the University of Iceland, the Faculty of Humanities and Social Sciences of the University of Zagreb, the National Library of Norway, and Machine Translation (MT) provider Iconic Translation Machines Ltd (now Language Weaver). The focus of the project was on providing data to improve the two Digital Service Infrastructures (DSIs) of eJustice and eProcurement, due to their strategic importance across the EU, in individual European Member States and in the associated countries of Iceland and Norway.

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 Connecting Europe Facility (CEF) projects such as iADAATPA (Castilho et al., 2019), ELRI[2] and Paracrawl.[3] This paper summarises the results from PRINCIPLE, focusing on its achievements, especially in terms of engagement with stakeholders and MT users, which promoted the continued collection of LRs with a view to improving and extending MT use.

2. Achievements

State-of-the-art domain-adapted neural MT (NMT) engines were built by the project partner Iconic for a number of early adopters (EAs) in all four countries. These public sector EAs included the Ministry of Foreign and European Affairs of the Republic of Croatia, the Icelandic Meteorological Office, the Icelandic Standards organisation, the Ministry of Foreign Affairs of Iceland, the Department of Justice in Ireland, Foras na Gaeilge, Rannóg an Aistriúcháin, the National University of Ireland, Galway, the Ministry of Foreign Affairs Norway, and Standards Norway. A small number of private companies also served as EAs on the project.

[2] https://www.elri-project.eu/
[3] http://paracrawl.eu/

These organizations collaborated with the project by sharing their LRs, and in return for contributing digital data sets to the project, they were offered dedicated state-of-the-art NMT systems. The development and subsequent evaluation of these MT systems according to the specific use-cases selected by the EAs served the purpose of validating the quality and demonstrating the actual value of the LRs collected by the project. Once the quality and effectiveness of the LRs had been verified, the data sets were shared with the wider community (subject to applicable licensing restrictions stipulated by the data providers) via ELRC-SHARE[4] and used to improve eTranslation.[5]

PRINCIPLE collected, validated and shared more than 50 data sets for the languages of the project. Most of these LRs are bilingual parallel corpora, but there are also a few monolingual and multilingual corpora, as well as glossaries. The project partners consistently ensured proper handling of copyright clearance and of issues related to intellectual property for LRs with all the relevant data providers. The majority of the LRs were contributed under the "CC-BY-4.0" licence, others under the "Open Under-Public Sector Information" and "Non-standard/Other Licence/Terms" licences, while a few remaining LRs were contributed under other miscellaneous licences. Some LRs contained proprietary and/or sensitive information and were therefore contributed exclusively to the Directorate General for Translation (DGT) of the European Commission to develop eTranslation, but they could not be shared with the general public. The majority of LRs are in plain text and TMX format, while some are in text with tab-separated values and text in comma-separated values, which ensures wider reusability and interoperability to benefit the largest possible number of users and applications.

In keeping with the aim of demonstrating the value of LRs being collected in the PRINCIPLE project for building MT systems, and demonstrating the benefits of MT especially to public sector users, an extensive MT evaluation was undertaken. This included an automatic evaluation using a range of metrics on both baseline and domain-specific systems and compared to a range of publicly available engines, as well as extensive evaluations conducted directly by public sector users. User evaluations included adequacy and fluency assessments, post-editing productivity, error analysis, and comparative systems rankings, all conducted by public sector translators independently of the project partners.

3. Conclusion

PRINCIPLE achieved its ambitious objectives, and the consortium partners worked successfully to collaborate with a range of existing and new data contributors in Croatia, Iceland, Ireland and Norway, so that valuable domain-specific LRs could be made available to the wider community. Our presentation at EAMT 2022 will give an overview of the main achievements of the PRINCIPLE project, with a focus on the set of public and private data holders and their use cases. In this context, we will discuss the range of LRs that have been gathered, and present an overview of the evaluation processes that were undertaken for the customised neural MT engines, with EAs in the four countries involved.

Acknowledgements: PRINCIPLE was 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.

[4] https://elrc-share.eu

[5] https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en