National Language Technology Platform (NLTP): The Final Stage

Artūrs Vasiļevskis[1], Jānis Ziediņš[2], Marko Tadić[3], Željka Motika[4], Mark Fishel[5], Bjarni Barkarson[6], Claudia Borg[7], Keith Aquilina[8] and Donatienne Spiteri[9]

[1] Tilde, Riga, Latvia
arturs.vasilevskis@tilde.com
[2] Culture Information Systems Centre, Riga, Latvia
janis.ziedins@kis.gov.lv
[3] University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb, Croatia
marko.tadic@ffzg.unizg.hr
[4] Central State Office for the Development of Digital Society, Zagreb, Croatia
Zeljka.Motika@rdd.hr
[5] University of Tartu, Tartu, Estonia
fishel@ut.ee
[6] Reykjavik University, Language and Voice Lab, Reykjavik, Iceland
bjarnibar@ru.is
[7] University of Malta, Valletta, Malta
claudia.borg@um.edu.mt
[8] Malta Information Technology Agency, Blata l-Bajda, Malta
keith.aquilina@gov.mt
[9] Office of the State Advocate, Valletta, Malta
donatienne.spiteri@stateadvocate.mt

Abstract. The final stage and the demo of the National Language Technology Platform (NLTP) developed within the CEF action of the same name is presented in this paper. The action aims at combining the most advanced language technology tools and solutions in a new state-of-the-art, artificial-intelligence-driven, web-based national platform for language technology oriented primarily towards users from public administrations of partner states. The Platform combines into a single framework the CAT tools, the TMs usage and management, the terminology management, several different MT engines and other language technology modules.

Keywords: machine translation, CAT tools, parallel corpora.

1 Introduction

The paper presents the final stage on the CEF action National Language Technology Platform (NLTP). The general aim of the action is to combine the most advanced Language Technology (LT) tools and solutions in a new state-of-the-art, artificial-intelligence-driven, web-based national platform for LT. The action is in its final stage with the fully functional systems being deployed at the level of partner states

Proceedings of the International Conference HiT-IT 2023, pages 203–208,
7-9 July 2023, Naples, Italy
https://doi.org/10.26615/issn.2683-0078.2023_019

(Latvia, Croatia, Estonia, Iceland, and Malta). In parallel, the planned data collection has been completed and consequently the machine translation (MT) systems training is also finalised. The paper is structured as follows: in Section 2 the previous projects and related work are presented. The targeted users are described in Section 3 while the details of the development process are given in Section 4. In Section 5 we provide the information about the NLTP sustainability and possible future directions.

2 Related work

The developed solution in NLTP[1] builds on the already existing `hugo.lv` platform and the results of the *EU Council Presidency Translator* (INEA/CEF/ICT/A2018/1762093)[2] action, which have proven beneficial over multiple years of active use. However, these two predecessors have been substantially extended into NLTP in order to provide public administrations, SMEs and the general public with secure access to high quality MT and integration with computer aided translation (CAT) tools, e-mail and web plug-ins etc., for translation of texts, documents and web pages. At this stage the set of offered services is considered final, but the modular design of the platform allows it to be enriched with additional LT services beyond this initial set. The NLTP modules and the overall structure of the platform are presented at Figure 1. The broader context and the introduction to NLTP can be consulted in two previous papers [2, 3] where motivation and overall view has been presented.

3 Users

In its final form NLTP is adapted, localised, and sustainably deployed by the public administration bodies in partner states, while its development is supported at the same time by local research institutions as complementary partners. In the case of Iceland and Estonia, the research partners were given the role of public authorities as well. Additionally, the NLTP was customised to the specific needs of public administrations so it provides translation using our own MT systems, but it is also additionally linked to eTranslation[3] services, thus enabling translations into and from the 24 official EU languages and other languages offered by eTranslation.

After the user needs were modelled following the estimated overall general needs, the additional specific requirements have been collected through a survey about LT needs and expectations in the public administration, that has been run in partner states. The example of such analysis of the survey in Croatia was presented in [1]. According to this survey e.g. 67% of users were familiar with CAT tools and 33% with MT, in 45% of institutions no LT is being used, the most useful LT in public administration are CAT tools (29%) and MT (11%).

[1] `https://nltp-info.eu`

[2] `https://presidencymt.eu`

[3] `https://ec.europa.eu/digital-building-blocks/wikis/display/CEFDIGITAL/eTranslation`

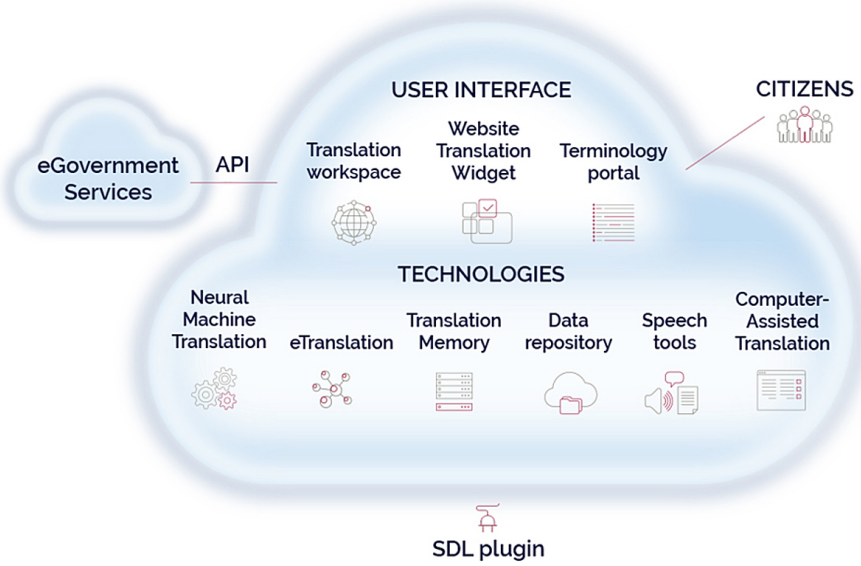

Fig. 1. The NLTP modules and overall structure.

The NLTP increases the efficiency of translation, the reuse of translation memories and the use of the existing high-quality MT technologies. Additionally, the action also integrates speech technologies for selected languages with automatic speech recognition and/or text-to-speech services. However, speech modules are not available for all languages at this moment since for some of them no such modules exist at all.

4 Development

Beside already customary services in the form of MT for pasted text and uploaded documents, the NLTP also provides additional valuable services such as a professional translation environment in the form of CAT tool integration through a simple-to-use online access in a plain browser and coupled with a number of other technological solutions, such as a translation widget, browser plugin, commercial CAT tool plugins, etc. This set of services ensures the widest possible reach to the users since these services could cover the needs of many users when they have to deal with the multilingual content. Providing public administration employees with a free, easy-to-use professional translation environment will further increase their productivity by creating a

cycle of use and reuse of translated content through translation memories (TM) accumulated in the process.

The platform also integrates services for terminology management linked with national terminology databases, as well as common European IATE[4] terminology database.

NLTP features MT systems tailored to the specific domains of administrations following their specific language, terminology, and communication styles. Examples of domains are legal, financial, medical and other areas of public administration that feature specific use of language. Customization maximizes translation quality for the local languages of the hosting country.

Each national variant of NLTP has been adapted according to the desired visual template, the interface and the help system have been localized in the national language, while for international users the interface in English is also available.

For each national variant a technical solution was packed in a Docker[5] installation and it has been integrated into the existing digital services at the national level.

4.1 Deployment

The platform was developed according to the common overall concept, but since the current eGovernment systems in partner states differ substantially, for each partner state a deployable variant had to be adapted to the needs of public administrations at the national level.

An example of such variant of deployment in Croatia can be seen in the Figure 2 and Figure 3. For instance, in Croatia the NLTP became an integral part of horizontal digital eGovernment services that are accessible by everyone working in the public administration at any level: national, regional and local. Also, this horizontal services are offered to anyone who has the authentication and authorisation privileges for lowest level of eGovernment services and this practically encompasses the whole public sector. There are also plans to offer these services to SMEs.

The similar deployment was conducted in other partner states, but adapted to their specific conditions and needs.

4.2 Additional datasets

Additionally, a number of domain-specific parallel datasets has been collected for five languages (Latvian, Croatian, Estonian, Icelandic, and Maltese) coupled with English. These datasets will be made available through the ELRC-SHARE[6] repository in the Translation Memory eXchange (TMX) or similar compatible format. Since the sources of data are predominantly coming from the public domain, the data will be accessible under permissive licences.

[4] https://iate.europa.eu
[5] https://www.docker.com
[6] https://elrc-share.eu

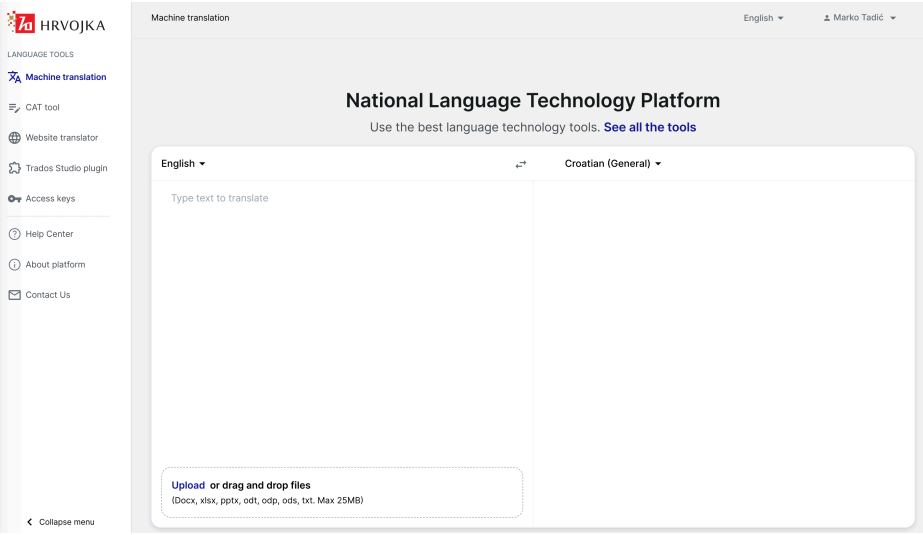

Fig. 2. Example of deployment in Croatia named Hrvojka, available at `https://hrvojka.gov.hr`. The typical MT service for translation of pasted text or uploaded document is presented with the web interface set to English for presentation clarity.

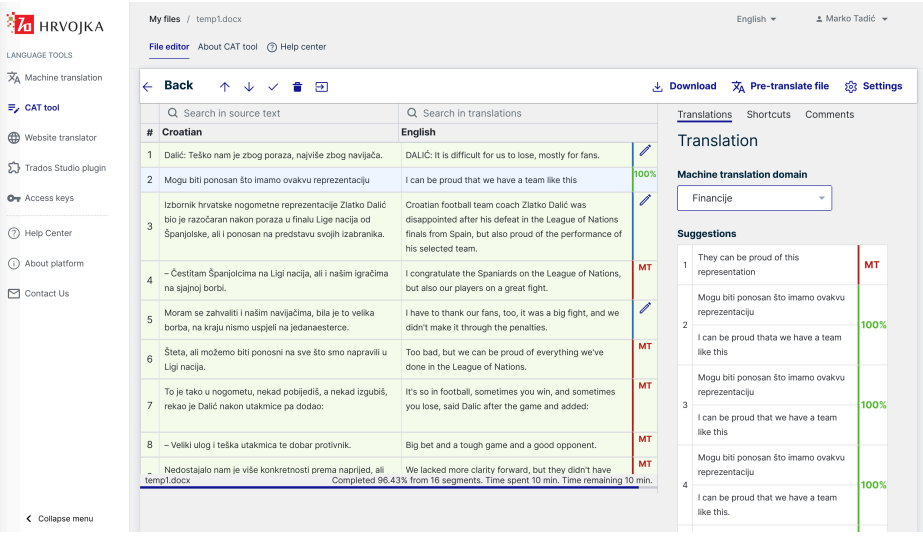

Fig. 3. Example of deployment in Croatia named Hrvojka, available at `https://hrvojka.gov.hr`. The CAT service running in a plain browser is presented with the web interface set to English for presentation clarity.

5 Sustainability and Future Directions

The public administration partner institutions are responsible for the sustainability of each national NLTP after the action ends. This has been provided by securing its inclusion into the national infrastructures for eGovernment as cloud services. This will enable multilingual access to and by public administrations, while, at the same time, the integration with public digital services offered in languages of the EU and EEA will be fostered.

For future research and development directions, similar platforms could be developed and deployed for other EU member states, and in this respect this action can be regarded as the proof-of-concept.

Also, with the introduction of new European language data sharing and processing initiatives such as Language Data Space and European Data Infrastructure Consortium being established for the language data, it is expected that services similar to NLTP would become more frequent and more readily available.

Acknowledgements

The work reported here was supported by the European Commission through the CEF Telecom Programme (Action No: 2020-EU-IA-0082, Grant Agreement No: INEA/CEF/ ICT/A2020/ 2278398, duration 2021-04-01–2023-06-30).

References

1. Motika, Ž., Didak Prekpalaj, T., Horvat Klemen, T., Košćec Perić, M.: Predstavljanje projekta *Nacionalna platforma za jezične tehnologije* (invited lecture). In: MIPRO 2022 - 45th Jubilee International Convention: CIS-AIS – Artificial Intelligence Systems, Opatija (2022), http://www.mipro.hr/MIPRO2022.CIS-AIS/ELink.aspx, last accessed 2023/06/19
2. Tadić, M., Farkaš, D., Filko, M., Vasiļevskis, A., Vasiļjevs, A., Ziediņš, J., Motika, Ž., Fishel, M., Loftsson, H., Guðnason, J., Borg, C., Cortis, K., Attard, J., Spiteri, D.: National Language Technology Platform for Public Administration. In: Aldabe, I., Altuna, B., Farwell, A., Rigau, G. (eds.) Proceedings of the LREC 2022 workshop Towards Digital Language Equality (TDLE 2022), pp. 46–51. European Language Resources Agency, Marseille (2022).
3. Vasiļevskis, A., Ziediņs, J., Tadić, M., Motika, Ž., Fishel, M., Loftsson, H., Guðnason, J., Borg, C., Cortis, K., Attard, J., Spiteri, D.: National Language Technology Platform (NLTP): overall view. In: Macken, L., Rufener, A., Van den Bogaert, J., Daems, J., Tezcan, A., Vanroy, B., Fonteyne, M., Barrault, L., Costa-Jussà, M., Kemp, E., Pilos, S., Declercq, C., Koponen, M., Forcada, M., Scarton, C., Moniz, H. (eds.) Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, pp. 343–344. European Association for Machine Translation, Ghent (2012).