Proceedings of the TDLE Workshop @LREC2022, pages 46–51
Marseille, 20 June 2022

National Language Technology Platform for Public Administration

Marko Tadić[1], Daša Farkaš[1], Matea Filko[1], Artūrs Vasiļevskis[2], Andrejs Vasiļjevs[2], Jānis Ziediņš[3], Željka Motika[4], Mark Fishel[5], Hrafn Loftsson[6], Jón Guðnason[6], Claudia Borg[7], Keith Cortis[8], Judie Attard[8], Donatienne Spiteri[9]

[1]University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb, Croatia,
{marko.tadic, dasa.farkas, matea.filko}@ffzg.hr
[2]Tilde, Riga, Latvia, {arturs.vasilevskis, andrejs.vasiljevs}@tilde.com
[3]Culture Information Systems Centre, Riga, Latvia, janis.ziedins@kis.gov.lv
[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, School of Technology, Reykjavik, Iceland, {hrafn, jg}@ru.is
[7]University of Malta, Valletta, Malta, claudia.borg@um.edu.mt
[8]Malta Information Technology Agency, Blata l-Bajda, Malta, {keith.cortis, judie.attard}@gov.mt
[9]Office of the State Advocate, Valletta, Malta, donatienne.spiteri@stateadvocate.mt

Abstract

This article presents the work in progress on the collaborative project of several European countries to develop National Language Technology Platform (NLTP). The project aims at combining the most advanced Language Technology tools and solutions in a new, state-of-the-art, Artificial Intelligence driven, National Language Technology Platform for five EU/EEA official and lower-resourced languages.

Keywords: machine translation, CAT tools, parallel corpora, National Language Technology Platform

1. Introduction

Multilingualism is one of Europe's fundamental values, alongside freedom of expression and freedom of movement. The European Charter of Fundamental Rights perpetuates the rights and freedoms of European citizens and ensures that linguistic diversity is maintained as a cornerstone of European policy. While the diversity of languages used in Europe is a true treasure, the barriers resulting from language insularity create big challenges for the cohesion of the European Union Digital Single Market[1,2] (DSM), public e-services, and public administrations. Many e-services provided by national public administrations are still available only in the official language of the respective country, which greatly restricts their access to guest workers, visitors, foreign investors and many other users who do not comprehend the local language. In case the e-services and public information are also provided in English and other languages, translation creates significant expenses and management burden.

In recent years, the European Union and member states have invested in various programmes to advance Language Technologies (LTs) as an efficient solution for breaking language barriers – from large language resoures collecting campaigns like ELRC[3], over the EC eTranslation services[4], up to establishing the common European marketplate for language technologies like ELG[5]. This has resulted in numerous LT tools, and services that have demonstrated their advantages across a wide range of national and international projects and have already proven their usefulness to public administrations. Nevertheless, their current use is limited because as individual solutions and technologies, they are used only by a particular target group, in a specific context, only in a few institutions and countries.

This paper presents the work in progress on the collaborative project of several European countries to develop *National Language Technology Platform (NLTP).*[6] The project aims to unite the most advanced LT tools and solutions developed in the CEF AT and other European and national programmes in a novel state-of-the-art, Artificial Intelligence (AI) driven software solution – NLTP. NLTP will provide national public administrations, SMEs and general public with mature, tightly integrated Machine Translation (MT) and other LT services (e.g. terminology management, translation memories, speech tools for selected languages, etc.) that will serve as an efficient way to enable multilingual access to information and public online services.

By providing essential LT support to various eGovernment services, NLTP is positioned to become an essential element of eGovernment infrastructure.

[1] https://ec.europa.eu/commission/sites/beta-political/files/dsm-factsheet_en.pdf [accessed 2020-05-16].

[2] https://ec.europa.eu/commission/presscorner/detail/en/IP_15_4919

[3] https://lr-coordination.eu

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

[5] https://www.european-language-grid.eu

[6] https://www.nltp-info.eu

The structure of the paper is as follows. We describe the composition of the consortium and represented languages in Section 2, followed by a brief description of the current state of the development of LT for each language in Section 3. In Section 4, the general goals and the collection of additional language resources is explained, while in Section 5 we conclude with a discussion on the targeted users of the NLTP and the user survey.

2. Consortium and Languages

The NLTP project is implemented as a CEF Telecom Action that includes partners from five countries: Latvia, Croatia, Estonia, Iceland and Malta, and it deals with LT for their respective languages: Latvian, Croatian, Estonian, Icelandic, Maltese. The consortium is deliberately composed of at least one partner (either academic or industrial) per country, that provides the expertise in LT for its language, and usually combined with one or more partners coming from public administration. In the case of Estonia and Iceland, the sole academic partners received the role of institution consulting the public administration. The project is initiated and coordinated by the LT company Tilde. The detailed list of partners can be found at the official web page of the project[7].

The five languages involved are all official languages of their respective countries, but at the same time they represent language communities with small to moderate number of speakers, starting in total with ca. 380,000 for Icelandic up to ca. 4,900,000 for Croatian, considering also citizens in other countries as emigrants of the first generation. The position of the official languages provides a unique opportunity to advocate for the state support of the usage of LT, and that is what is expected from the partners from public administration.

3. State of LT in ELE Language Reports

In this section we present in brief the state of the development of LT that was presented in more detail in the Language Reports as deliverables within the European Language Equality (ELE) project[8]. This state of development can serve as the baseline starting point, which could be used for measuring the project's progress by the end of its duration. Also, this illustrate the different stages of LT development and relevant gaps that exist between the consortium languages, as well as their different starting points. However, for all languages involved the need of more domain specific bilingual data and persistent national infrastructure for LT is stressed. We strongly believe that NLTP could contribute to both urgent needs.

3.1 Latvia

Since 2012, when the META-NET white papers were published (Skadina et al., 2012), significant progress has been made in the development and deployment of different language resources and tools for the Latvian language. Latvian also has good support in terms of more advanced technologies, such as MT, as well as speech recognition and synthesis. At the same time, solutions involving state-of-the-art natural language understanding are not so developed. There are gaps regarding the availability, size, and technology readiness level of language resources. More models, tools, as well as computational, human, and financial resources would benefit the current state of LT for Latvian. Significant gaps were identified in monolingual and multilingual data – written, spoken, and multimodal, especially when it comes to open-data or open-access language resources. Fine-tuning domain-specific engines is also more complicated than desirable, as the current domain-specific data is insufficient and lacks substantial open-access monolingual text corpora. Overall, the current LT situation in Latvia is fragmented, but going in the positive direction and continuously improving while most areas are considered to have “fragmentary support”. In the whitepaper, it was mentioned that dedicated long-term LT programs would benefit research and industrial activities, with NLTP addressing the latter.

3.2 Croatia

Technological support for Croatian has progressed in a number of LT areas compared to the state of affairs described in the META-NET White Paper (Tadić et al., 2012). Digital language resources have both increased in number and volume while they also improved in quality and variety. Resources, basic NLP tools and LT services are provided by academia, research institutes and occasionally private companies as outputs of various research projects, usually coordinated by academic institutions, predominantly funded by EU or national funds, and rarely self-funded. Some significant progress has been made with respect to available corpora and lexica, language models, text processing tools, and MT, while there is still a serious underdevelopment in the field of speech processing (both synthesis and recognition). The available datasets originate from a variety of sources and they cover several thematic domains, text types; they are available as raw or annotated; and come as monolingual, bilingual or multilingual resources. However, their individual size is lagging behind in terms of appropriateness for building large language models or robust, ready to use tools and applications.

One of the long-term intentions is to secure the presence of Croatian NLP modules in the major NLP platforms (commercial and non-commercial) such as spaCy, FreeLing, NLP Cube, TextRazor, Cloud Natural Language, Apache Open NLP, etc., in order to secure the sustainability and wider usage of LT for Croatian and, consequently, its digital language equality with other languages. The inclusion of Croatian in NLTP will certainly add to this goal.

3.3 Estonia

During the last decade some LT fields have advanced significantly and for Estonian there are better and bigger corpora of contemporary written language and bigger treebanks, but several gaps that were identified by the Meta-Net White Paper in 2012 (Liin et al., 2012) are still there: text generation is still under-developed and we lack annotated semantic resources and tools for semantics. The existing tools cover the basics of text analysis – sentence segmentation, tokenisation, morphological analysis,

[7] https://www.nltp-info.eu

[8] https://european-language-equality.eu

syntactic parsing – for standard written language. The overall quality of MT and especially of speech technologies is also quite satisfactory, but only for standard language. However, Estonian lacks both annotated data and tools for certain tasks and, as annotating data is a time- and workforce-consuming process, it can be seen as an even bigger obstacle. There are large web-crawled corpora for Estonian, but less domain-specific corpora or, if such resources exist, they are not publicly available. Accordingly, resources need to be made available – as has been done successfully in other countries – to persuade Estonian data-holders of the benefits of sharing such data sets.

Most of the work in Estonian LT is done at academic institutions and is project-based. Once the project is over, the developed resources are not updated any more. Accordingly, there is a need for an infrastructure for keeping these models and tools up-to-date once the project has ended so that Estonia can continuously benefit from that important work. The national Estonian CLARIN consortium and its participation in CLARIN ERIC will strenghten this role in future.

3.4 Iceland

Ten years ago, the META-NET White Paper Series described a serious situation for the Icelandic language. At that time, Icelandic was one of the four European languages that fell into the "weak/no support" category regarding the level of support for speech processing, MT, text analysis, and speech and text resources (Rögnvaldsson et al., 2012). This alarming situation triggered concerns and discussions among politicians in the following years, which eventually resulted in the establishment and financing of the Language Technology Programme for Icelandic (LTPI), 2019-2023 (Rögnvaldsson, 2020).

The goal of the LTPI is to make Icelandic usable in information and communication technology, by developing open-source language resources and tools. The LPTI consists of six core projects: Language Resources, NLP tools, Automatic Speech Recognition, Speech synthesis, Spell and Grammar Checking, and MT (Nikulásdóttir, 2020). Even though the LTPI is still being implemented, it has already made a very significant change to the level of language technology support for the Icelandic language, as is evident by the comprehensive listing of the deliverables found in Nikulásdóttir et al. (2022), and the CLARIN-IS repository. Nevertheless, after the LTPI ends, Icelandic will still lack various NLP tools and resources, which indeed shows the need for a continued governmental support of the LTPI (Rögnvaldsson, 2020).

The inclusion of Icelandic in NLTP fits well with the LTPI, particularly regarding the MT, Speech Recognition, and Speech synthesis core packages.

3.5 Malta

The main gaps in the development of LT for Maltese are present in three general areas: (i) tools (ii) resources and (iii) support.

As far as tools are concerned, Maltese still lacks the barest minimum required for a BLARK (Basic Language Resource Kit) as defined in (Krauwer,1998). So Maltese needs not only a solid set of building blocks that will serve to build more advanced applications, but also ready and universal access to them with the help of platforms like ELG[9]. The potential for MT is beginning to be appreciated thanks to the efforts by the EC, but more effort is required locally, beyond the limited timeframe of the NLTP project, for the necessary quality to be achieved in all the domains where it can usefully serve. This requires a concerted policy to facilitate the extraction and refinement of bilingual resources at their point of creation, i.e. public administration. Speech technology, and particularly Automatic Speech Recognition (ASR), is another priority, since it imparts a highly tangible quality to LT that makes sense to ordinary people in everyday situations. Finally, for Maltese there is also a complete lack of multimodal tools.

These days, almost all of the above tools are driven by machine learning, and thus their quality depends on the availability of suitable language resources. If intelligent multimodal and multilingual machinery that involves Maltese is expected, then appropriate multimodal and multilingual corpora are needed to train it. This requires a significant effort, which, if the resources are to faithfully reflect their inherent regional characteristics, must be developed at national level.

This evokes the third gap: support. LT has not received the kind of recognition that is normally afforded to language by various national institutions. If LT for Maltese is to thrive, it needs to be recognised as a national area of priority that requires nurturing, management and support. Most of the language resources and tools that exist today have been created opportunistically. This is a short-term expedient that creates gaps and discontinuities. Language resources and tools need to be commissioned to fit carefully identified needs, and curated on a permanent basis. This requires commitment at national level, and a serious budget, as seen in Spain, Estonia, The Netherlands and even smaller countries like Iceland (Nikulásdóttir et al., 2020). Currently, the institutions that are responsible for the Maltese language adopt a helpful stance towards LT, but have not really taken on board the commitment that is required to ensure that it flourishes and exploits its full potential.

4. NLTP Goals and Resources

In this section, we first state the general goals of NLTP, then describe the work that precedes the NLTP project and provides the background knowledge and results on which we build upon. In continuation, we describe the NLTP components and the specific needs of collecting additional language resources, with the emphasis on the parallel corpora for these five languages, that are particularly deficient with this type of language resources.

[9] https://www.european-language-grid.eu

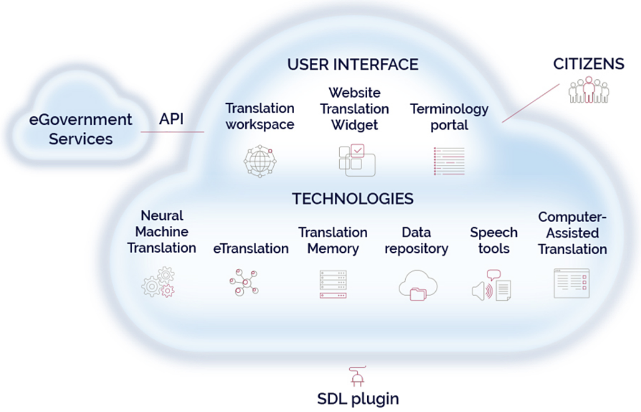

Figure 1. Schematic representation of NLTP components in the initial configuration.

4.1 General Goals

The aim of the project is to create a generic, customizable LT platform that any European country can easily adapt and deploy in their eGovernment infrastructure. The project partners are assessing the particular needs of their public administrations, tailoring the platform for their needs and deploying and integrating it into infrastructure of public digital services.

4.2 Related Work

The proposed solution for national LT platforms will rely on the existing *hugo.lv* platform[10], but it will also include the results obtained from the CEF action *EU Council Presidency Translator* (INEA/CEF/ICT/A2018/1762093)[11]. Both projects featured online MT services tailored for translation between the national language(s) and English in both directions. While the hugo.lv was oriented towards Latvian and English for use by Latvian public administration, the EU Council Presidency Translator covered seven languages in addition to English: Latvian, Estonian, Bulgarian, German, Romanian, Finnish and Croatian (chronologically ordered).

4.3 NLTP Components and Approach

The existing platform(s) will be substantially expanded into the planned NLTP in order to provide public administrations and the general public with a secure access to high quality MT and integration with Computer Aided Translation (CAT) tools, e-mail plug-ins, web-pages translation widgets etc., for translation of texts and documents (see Figure 1).

For both public administration and the public, this platform will be a convenient and secure environment for translating texts, documents and websites using fast and efficient newest generation Neural MT (NMT) technologies. Users will be able not only to translate entire documents in multiple languages preserving their formatting, but they can also review and edit the MT output, create and check terminology. Input and output will be either in text or voice form depending on the NLTP service provided by the consortium partners and selected by the user.

The NLTP will be adapted, localised, and sustainably deployed by the public authority bodies in the partner states as a part of the national eGovernment services and infrastructures, while its development will be supported by local research institutions as complementary partners. The inclusion of NLTP in national eGovernment infrastructures will secure the sustainability of the platform after the EU funding period.

Additionally, the NLTP will be customisable to the specific needs of each public administration. Adaptation will be carried out by redesigning the initial configuration with adapting the existing solutions developed within other projects, creating a unified and customizable user interface, and furnishing the platform with the latest neural MT (NMT) systems. NMT systems will be tailored to the specific domains of administrations using specific domain language, terminology, and communication styles. This can include legal, financial or other domains heavily used in public administration. Customization will maximize translation quality for the local languages of the hosting country.

The modular design of the NLTP allows the inclusion of other LT services when public administration express their needs for them, e.g. Automatic Speech Recognition for transcription of recorded sessions in order to produce meeting minutes, or Named Entities Recognition module in text preprocesssing step.

NLTP will be further linked to EC eTranslation services, thus enabling translations into and from the 24 official EU languages and other languages of the Digital Single Market. Equally, this platform is open for integration with other MT providers by simple integration using Docker container technology or open API connection.

The NLTP will facilitate the use of a professional translation environment with integrated terminology databases, CAT tools and (N)MT, all wrapped up in simple-to-use HTML front end and coupled with number of other technological solutions, such as a translation widget, browser plugin, commercial CAT tool plugins, etc.

4.4 Additional Resources

To customise MT systems of the platform for the domains most needed by national public administrations, a number of domain specific parallel data is being collected in all available digital formats (predominantly HTML and PDF). The processing is being done by several partners in order to reach the final targeted Translation Memory eXchange (TMX) format. The processing includes the boilerplate removal and/or PDF text extraction, automatic sentence splitting and sentence-alignment that will be checked for possible alignment errors. The bilingual language resources will be made available through the ELRC-SHARE[12] repository and other repositories. Since the sources of data are predominantly expected to come

[10] https://hugo.lv/en

[11] https://www.tilde.com/products-and-services/machine-translation/de-presidency

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

from the public domain, the data will be made accessible under permissive licences. The parallel corpora collection process will be described in a separate paper when the collection process will be completed.

5. Users

The NLTP is targeted to several target groups which are discussed in this section. In order to understand the public opinion towards the usage of LT in five countries, we have organised user surveys that are currently being run.

5.1 Targeted Users and User Requirements

The relevant stakeholders, which are expected as end-users of the platform are different bodies of public administration (local/regional/national), private sector (particularly SMEs), academia and citizens. NLTP specifications will be finally defined in collaboration with the consortium partners and relevant stakeholders using information collected in user surveys. The user groups are primarily defined through the user stories, in particular targeting user experience, i.e., defining the end-user, purpose and use of the country specific NLTP configuration. The user stories were initially defined at the project start, then updated/added accordingly during the NLTP development. The methodology for definition of user stories is based on (i) user identification and classification into group profiles, in order to place user requirements in a problem-driven context; (ii) using mock-ups of the system and its main services as a trigger in interactions with users.

The requirements are being analysed and prioritized, and the analysis will come up with the description of different types of services from the users' point of view. The requirements analysis will feed the functional specifications of the NLTP and subsequently drive the design of the architecture and its backend components.

5.2 User Survey

Targeted surveys are being run in the NLTP consortium member states with the intent to gain insight into current use, current needs, and potential future needs of the NLTP end-users and different public institutions which are adopting the NLTP concept. The surveys are being disseminated by the consortium representatives of each country through established network of contacts and responses are collected from public institutions as possible adopters and end-users. The data will be processed to identify gaps in language tools that the NLTP could fill, as well as other features that institutions in each country might benefit from. The survey results will also be used for preparation of national seminars and workshops that will promote the benefits of frequent LT usage.

6. Conclusions

We presented the NLTP, a project which aims to build a National Language Technology Platform in five EU/EEA countries intended for use primarily by public administration. The platform will initially offer what is needed the most by this category of users: the access to NMT systems (proprietary and eTranslation-based), to CALL environment in simple HTML interface, to terminology and translation memory management, but also to other LT services due to the platform's modular design.

Acknowledgements

The work reported here was supported by the European Commission in the CEF Telecom Programme (Action No: 2020-EU-IA-0082, Grant Agreement No: INEA/CEF/ ICT/A2020/2278398).

References

Krauwer, S. (1998) ELSNET and ELRA: Common past, common future, *ELRA Newsletter*, Vol 3 no 2. [http://www.elsnet.org/dox/blark.html]

Liin, K., Muischnek, K., Müürisep, K. and Vider, K. (2012) *Eesti keel digiajastul – The Estonian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in Digital Age. Springer, Heidelberg, New York, Dordrecht, London. [http://www.meta-net.eu/whitepapers/volumes/estonian].

Motika, Ž., Didak Prekpalaj, T., Horvat Klemen, T., Košćec Perić, M. (in press). Predstavljanje projekta "Nacionalna platforma za jezične tehnologije" In: *Proceedings of the μPro2022 Conference*, Opatija, May 2022.

Muischnek, K. (2022) *D1.12: Report on the Estonian Language*. European Language Equality. [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_12__Language_Report_Estonian_.pdf]

Nikulásdóttir A. B., Arnardóttir, Þ., Barkarson, S., Guðnason J., Gunnarsson, Þ. D., Ingason, A. K., Jónsson, H. P., Loftsson, H., Óladóttir, H., Rögnvaldsson, E., Sigurðsson, E. F., Sigurgeirsson, A. F., Snæbjarnarson, V., Steingrímsson, S., and Örnólfsson, G. T. (2020) Help Yourself from the Buffet: National Language Technology Infrastructure Initiative on CLARIN-IS. In *Selected Papers from the CLARIN Annual Conference*.

Nikulásdóttir A. B., Guðnason J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. (2020) Language Technology Programme for Icelandic 2019-2023. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*.

Rögnvaldsson, E., Jóhannsdóttir, K. M., Helgadóttir, S., and Steingrímsson, S. (2012) *Íslensk tunga á stafrænni öld - The Icelandic Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London. [http://www.meta-net.eu/whitepapers/volumes/icelandic]

Rögnvaldsson, E. (2022) *D1.19: Report on the Icelandic Language*. European Language Equality. [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_19__Language_Report_Icelandic_.pdf]

Rosner, M., Joachimsen, J. (2012) *The Maltese Language in the Digital Age / Il-Lingwa Maltija Fl-Era Diġitali*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London. [http://www.meta-net.eu/whitepapers/e-book/maltese.pdf]

Rosner, M., Borg, C. (2022) *D1.25: Report on the Maltese Language*. European Language Equality. [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_25__Language_Report_Maltese_.pdf]

Skadiņa, I., Veisbergs, A., Vasiļjevs, A., Gornostaja, T., Keiša, I., Rudzīte, A. (2012) *The Latvian Language in the Digital Age / Latviešu valoda digitālajā laikmetā*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London. [http://www.meta-net.eu/whitepapers/e-book/latvian.pdf]

Skadiņa, I., Auziņa, I., Valkovska, B., Grūzītis, N. (2022) *D1.22: Report on the Latvian Language*. European Language Equality. [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_22__Language_Report_Latvian_.pdf]

Skadiņš, R., Pinnis, M., Vasiļevskis, A., Vasiļjevs, A., Šics, V., Rozis, R., & Lagzdiņš, A. (2020). Language Technology Platform for Public Administration. In *Human Language Technologies–The Baltic Perspective* (pp. 182-190). IOS Press.

Tadić, M., Brozović-Rončević, D., Kapetanović, A. (2012). *Hrvatski Jezik u Digitalnom Dobu – The Croatian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London. [http://www.meta- net.eu/whitepapers/ volumes/Croatian].

Tadić, M. (2022) *D1.7: Report on the Croatian Language. European Language Equality*. [https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_7__Language_Report_Croatian_.pdf]