Human Language Technologies – The Baltic Perspective
I. Skadiņa and R. Rozis (Eds.)

doi:10.3233/978-1-61499-701-6-20

European Platform for the Multilingual Digital Single Market: Conceptual Proposal

Andrejs VASIĻJEVS[a,1], Jan HAJIC[b], Jochen HUMMEL[c], Josef van GENABITH[d] and Rihards KALNIŅŠ[a]
[a] *Tilde, Latvia*
[b] *Charles University in Prague, Czech Republic*
[c] *ESTeam, Sweden*
[d] *DFKI, Germany*

Abstract. This paper proposes a large scale initiative for the creation of the European Platform of online multilingual services to cover the multilingual needs of the Digital Single Market. We describe the three layers of the Platform: a solutions layer, an infrastructure layer, and a research layer. The infrastructure layer will combine mature language technologies in four clouds: Automated Translation Cloud, Human-Computer Interaction Cloud, Multilingual Knowledge Management Cloud, and European Language Cloud, encompassing basic services and language resources. We identify the key gaps and recommend targeted research activities to provide equal technology coverage for all EU languages.

Keywords. Language infrastructure, digital single market, automated translation, human-computer interaction, multilingual knowledge management, language resources

1. Introduction

This paper proposes a large scale initiative for the creation of the European Platform of online multilingual services to cover the multilingual needs of the Digital Single Market (DSM)[2]. It is based on the proposals[3] that the authors submitted to the European Commission after the Riga Summit 2015 on Multilingual DSM[4].

The European Commission has defined a Digital Single Market as one in which the free movement of goods, persons, services, and capital is ensured and where individuals and businesses can seamlessly access and exercise online activities under conditions of fair competition and a high level of consumer and personal data protection, irrespective of their nationality or place of residence [1]. To establish such a market, it is necessary to remove barriers and obstacles preventing equal access and use of digital services for all Europeans.

Although linguistic diversity is a fundamental value of Europe, the language barriers created by more than 80 languages (24 of which are official EU languages) create a major barrier for achieving the vision of DSM. According to a Eurobarometer survey [2], just

[1] Corresponding Author: Andrejs Vasiļjevs, e-mail: andrejs@tilde.lv
[2] http://ec.europa.eu/priorities/digital-single-market/
[3] http://www.lt-innovate.org/sites/default/files/Multilingual%20Platform%20Concept.pdf
[4] http://rigasummit2015.eu

over half of Europeans (54%) are able to hold a conversation in at least one additional language besides their mother tongue, and only 38% of Europeans speak English well enough to hold a conversation. Another Eurobarometer survey [3] shows that 59% of internet users only use their native language when writing emails, sending messages, or posting comments on the Web, while 42% of Europeans do not buy products or services online if they are not in their native language.

Language technologies are the key to crossing language barriers in Europe and creating a truly multilingual DSM. Drawing on years of innovative research and development, much of it supported by EU funding, the European industry has developed multiple mature technologies to enable multilingualism.

At the same time, industry offerings are very fragmented as companies either focus on only larger languages or on their local markets. Many EU languages are disadvantaged due to gaps in technology coverage or poor quality in comparison with a handful of larger languages for which much better solutions are available. Importantly, however, it is not just the size of the languages (e.g., measured in terms of the number of native speakers) that correlates with the quality of services: some of the larger languages also exhibit linguistic properties such as rich morphology and/or less constrained word order, which make it harder for machines to process them. This is demonstrated by the comparative study of META-NET [4]. After assessing the level of support through language technology for European languages, the experts concluded that digital support for 21 of the 31 languages investigated is weak and doesn't match the needs of the digital age. As a result, many European companies and citizens are basically excluded from future key innovations.

If the fragmented industry cannot deliver multilingual capabilities for all member states, the public sector should make a difference by kick-starting the creation of the necessary infrastructure.

In order to create a truly multilingual DSM, we propose to create a **European Platform for the Multilingual DSM** (Multilingual Platform). This Multilingual Platform will combine mature language technologies (LT) in several clouds and provide services that will be made available to startups, SMEs, IT integrators, industry, and the public sector.

The Platform will encompass three layers: solutions (Layer 1), infrastructure (Layer 2), and research (Layer 3), corresponding to version 0.5 of the Strategic Agenda for the Multilingual Digital Single Market document (Strategic Agenda) [5].

2. Innovative Solutions (Layer I)

Layer 1 combines language technology solutions that are built by SMEs using the services available in the Multilingual Platform. Solutions will enable e-commerce and digital services providers, public administrations, cross-border public service providers, and other stakeholders to easily integrate multilingual capabilities in their daily work. These solutions can also be integrated into the workflows of large-scale organizations – such as EU institutions, NGOs, media outlets, and corporations – and also be made available to the public, enabling multilingualism on a pan-European scale.

Enabled by the basic language technology services of the infrastructure clouds (Layer II), LT SMEs and solution providers will create components to cover a range of client and market specific needs. Fragmentation of these commercial offerings will be

overcome by the establishment of the LTI Cloud - the LT industry's one-stop marketplace for LT components.

LTI Cloud aggregates commercial LT components in order to make it easier for solution providers to discover, test, integrate, and license language technologies. The LTI Cloud is a SaaS wrapper around language technology (LT) components and functions as a marketplace. It will make it easy for start-ups, IT departments, system integrators, and software companies to plug 'n' play language technology. LT companies can market their capabilities, build LT components based on others, and use the LTI Cloud as a customer acquisition channel.

A pilot implementation of the LTI Cloud[5] is already created in the framework of FP7 CSA project MLi[6].

3. Language Service Infrastructure (Layer II)

The infrastructure layer will combine the mature language technologies in four distinct clouds: *Automated Translation Cloud, Human-Computer Interaction (HCI) Cloud, Multilingual Knowledge Cloud,* and *European Language Cloud*. The clouds are interoperable and based on a core foundation of language resources. These infrastructures must not compete with the software industry or service providers. They provide fundamental facilities and systems that private companies cannot build efficiently or profitably.

The governance of the infrastructure layer should involve a broad coalition of stakeholders made up of representatives from industry associations and research organizations (i.e., LT-Innovate, META-NET, CLARIN, ELRA, etc.), European institutions (i.e., DG Connect, DG Translation, etc.), and key user groups (e.g., BDVA). It will serve as the main partner of the European Commission in guiding implementation and operations of the infrastructure.

3.1. Automated Translation Cloud

As one of the most critical language services, translation provides access to information and digital services across languages. The machine translation services of global providers (e.g., Google Translate, Bing Translator) are widely used, but they are not able to provide high quality translation for all European languages and to meet the specific requirements of the high demands of public and private applications, such as confidentiality and adaption to different domains.

The Automated Translation Cloud will combine the highest-quality machine translation services for each EU language and all language pairs, provided by CEF Automated Translation[7], EU Member States, and commercial providers. These MT services will be available in multiple domains. The services will enable solution developers to integrate instant translation capabilities into any platform or application, including mobile apps, web portals, or e-commerce sites (see Solutions Layer I).

- **CEF.AT services:** MT services provided by CEF.AT for the CEF DSIs and EU public administrations.

[5] http://www.lticloud.eu
[6] http://mli-project.eu
[7] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

- **National MT services:** MT services developed by EU Member States for use in national public sectors (e.g., Hugo.lv [6], Versti.eu).
- **Commercial services:** MT services developed by commercial providers (e.g., LT-Innovate and TAUS members).

3.2. Human-Computer Interaction (HCI) Cloud

Proliferation of ubiquitous devices like tablets and mobile phones, and increasingly also smart appliances and robots, requires efficient human-computer interaction beyond traditional graphical and text based interfaces. Voice commands are quick and intuitive, allowing to control devices in a natural way. Though the latest research has yielded strong results in voice based interaction (e.g., Siri and Cortana), this success is confined to larger languages. Speakers of smaller languages do not have access to technology that would allow them to interact with devices in their own language.

The HCI Cloud will address this issue by providing speech and other human-computer interaction services for all EU languages, which can then be used to build robust multilingual solutions.

The HCI Cloud combines speech services – automatic speech recognition (ASR), text-to-speech (TTS), dialog management (DM), speaker and language identification, keyword spotting, voice search, audio and video indexing using phonetic and word level indexes, dysfluency detection and removal, specific tools for connecting automatic translation and speech processing for spoken text translation tools, and multimedia communication modules for basic functionality in sign language recognition and generation, image search, image and video object recognition and tracking, and text-in-image and -video recognition – for all EU languages.

The Human-Computer Interaction Cloud will be based on mostly Open Source tools, such as the KALDI speech recognition generator based on state-of-the-art machine learning methods, which will in turn utilize the language resources collected and annotated. Similar technology exists for the other tasks, such as text-in-image identification and object detection in images and video.

3.3. Multilingual Knowledge Cloud

The Multilingual Knowledge Cloud combines semantic interoperability services for making e-Government and commercial services interoperable and enabling knowledge based data processing. It is essential for the implementation of the European Interoperability Framework (EIF), which is planned to become mandatory for all public IT projects. Semantic Interoperability required by the EIF is achieved by deploying multilingual meaning and knowledge assets. These assets will be pooled and exposed in the Multilingual Knowledge Cloud. It will enable meanings to be carried across language boundaries via data structures and data elements that are specific to different sectors. The Multilingual Knowledge Cloud will embrace existing developments at EU institutions. For example, the European Office for Harmonization created the multilingual knowledge system TMClass in order to be able to process Community Trade Mark applications in all official languages.

ISA (Interoperability Solutions for European Public Administrations) can provide the standards and processing know-how for aggregating Semantic Interoperability Assets, making them discoverable in a central network and promoting their usage, as well as filling-up of missing languages. It would define and host the access layer so that larger

IT companies, SMEs, and start-ups will be able to build on these assets and flourish by combining them with other types of data to build and market new applications.

3.4. European Language Cloud (ELC)

All language processing applications (search, mining, writing, speech, translation, etc.) depend on a basic natural language processing (NLP) infrastructure. The European Language Cloud (ELC) is a public infrastructure that provides the basic functionality required to process unstructured content. Through an API, it provides basic language technology services such as tokenization, stemming, morphology analysis, part of speech tagging, named entity detection, identification of measurements, currencies, formulas, etc. for all languages, in the same base quality, under the same favorable terms. National institutions in charge of language maintenance provide data and standards. The ELC builds on language resources and forms the basis for all LT efforts in text and speech processing.

3.5. European Language Resources

At the core of the various clouds making up the platform is a repository of language resources. These resources include language data such as parallel and monolingual texts, lexicons, and terminologies used for building MT services, voice recordings used to build ASR services, and monolingual texts used for developing linguistic tools. Also included are the language resources provided by CEF.AT. The resources will be provided by the European language research community, as well as by various EU Member States through the efforts of pan-European language-data collection initiatives.

The European Language Cloud should be based on the existing repositories of ELRA [8] and META-SHARE [9] [7], providing sustainable support, coordination, and funding mechanisms.

4. Research (Layer III)

Layer 3 addresses the gaps in coverage for all EU languages and provides novel methods to improve the quality and applicability of language technologies. This layer is described in detail in the Strategic Agenda.

In order to identify the gaps in language technology, data (resources), and coverage with respect to relevant European languages, we rely heavily on the META-NET Language White Papers [4].

4.1. Language Resource Gaps

Data is the fuel of most modern state-of-the-art language technologies. Insufficient amounts of training data limit the quality of the language technology system in question. This has serious consequences for applications: Data Value Chains may end up locked into language silos. Out of the 31 European languages covered by the Language White

[8] http://catalog.elra.info/
[9] http://www.meta-share.org/

Papers [4], 21 have fragmentary or low language resource support; only English is considered as having "very good support".

4.2. European Language Cloud Gaps

BLARK [8] elements for 28 languages are missing or of insufficient quality, mainly due to small or no data resources available for a particular element. There is a lack of lexical resources for knowledge-based approaches, esp. at the semantic level (typical examples: WordNet, FrameNet, propositional and valency dictionaries).

Recommendation: Identify existing resources (reviewing META-SHARE, CLARIN VLO and national nodes, traditional data marketplaces like ELRA and LDC) and how they fit the current state-of-the-art language-universal or near-universal open technologies; then fill the gaps found by resource collection or creation and coupling with state-of-the-art open technologies to create at least comparable-quality BLARKs for all languages (in several steps/batches).

4.3. Automated Translation Cloud Gaps

- Lack of high-quality MT system pairs due to lack of parallel and monolingual data for MT systems training in 28 languages (in various amounts and sizes fitting the language types and families)
- Technology for more efficient generalization and training off less data than current technologies
- Technologies for fast user feedback incorporation

Recommendation: Extend and continue data collection efforts (such as those started by ELRC [10] under the CEF programme), especially from public and other aforementioned domains, data clearance (IPR) efforts, and cooperation with data holders to clear and/or open more data for general use (TAUS, ELRA, LDC, national sources, public administrations), with focus on the 28 under-resourced languages. Support research activities for new technology development under research and innovation actions and possibly also FET (advanced ML methods, DNNs, etc.). Support methods for more efficient data collection.

4.4. Automated Translation Cloud Gaps

- Multilingual linked open data, and/or LOD linked to language resources, focusing on leveling the support for the 25 lagging languages (example: the medical UMLS ontology is available only in Basque, Croatian, Czech, Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Latvian, Norwegian, Polish, Portuguese, Spanish, and Swedish, i.e., 14 languages are missing, and source coverage is much smaller even for those included when compared to English)

[10] http://lr-coordination.eu

- Language resources for development of robust key technological components (text classification, topic modeling, entity linking and normalization, event detection, textual entailment)

Recommendation: Start prioritized data collection efforts in the two areas identified above. Support research in alternative methods for using unannotated and/or smaller amounts of data to achieve the same goals under research, innovation, and e-Infrastructure areas.

4.5. Human-Computer Interaction Cloud Gaps

- Basic language speech data, especially for the 22 languages with fragmentary to weak/no support
- Robust ASR technology for adverse conditions
- Dialog management technology with continuous learning capability, including available conversation recordings in all languages and major domains of interest (healthcare, home/vehicle automation, public services, and others as identified by the DVC)
- Technology for combining modalities, including combined datasets for all languages

Recommendation: On top of the speech data collection effort, which can start immediately, we recommend to create a conversational and multimodal data collection platform within the Language Resources cloud. This would enable collection of conversation data, especially over the telephone, for the selected domains.

We strongly recommend funding cutting edge research on modality combination (speech, text, and vision) under research and innovation projects under the Data Value Chain and FET, so that the CEF scheme may be used for the HCI cloud in two-to-three years' time and so that data will be available for broad and inclusive language coverage.

5. Summary Conclusions

Concerted large scale actions are needed to break the language barriers for the Digital Single Market. The chart below illustrates the above described platform for enabling a Multilingual Digital Market. The proposed platform links together all the necessary elements to enable essential language technologies for all European languages that will serve the needs of European public administrations, the private sector, and citizens. This approach will consolidate fragmented activities in current EU programmes Horizon 2020 and Connecting European Facilities, national programmes in EU member states, and various initiatives of research and industry communities.

Acknowledgements

The work described herein has been supported in part by the CRACKER H2020 project of the EC, project No. 645357, in part by the Research infrastructure LINDAT/CLARIN,

project No. LM2015071 of the MEYS CR, and partly by the FP7 project MLi, project No. 610951.

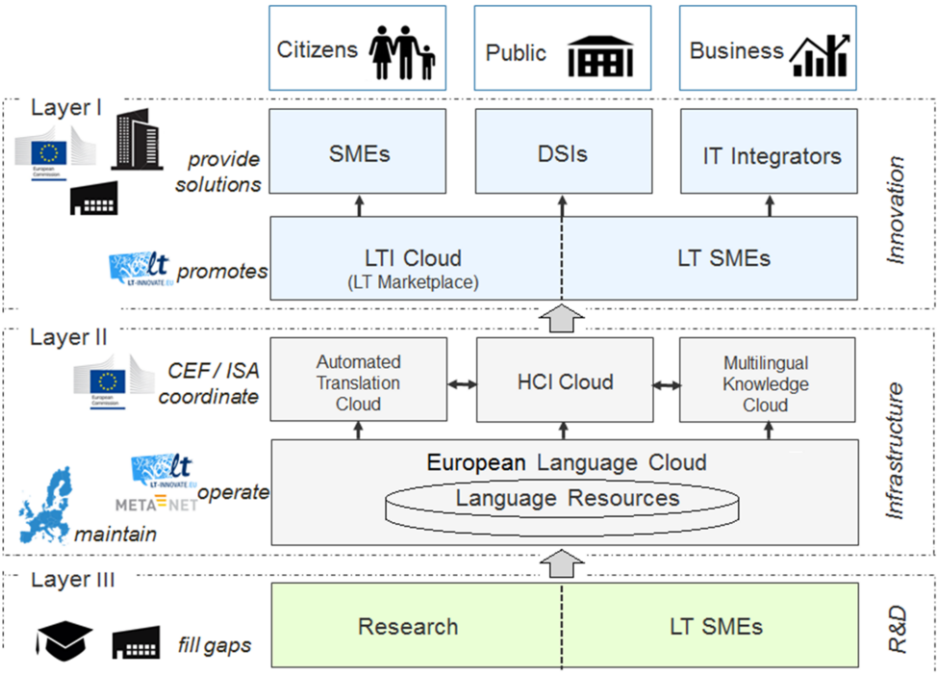

Figure 1. Diagram of the European Platform for the Multilingual DSM.

References

[1] A Digital Single Market Strategy for Europe. Communication from the Commission to European Parliament, the Council, the European Economic and Social Committee and Committee of the Regions. http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52015DC0192&from=EN, Brussels, 2015

[2] Europeans and their Languages. Special Eurobarometer 386, 2012.

[3] User language preferences online. Flash Eurobarometer, 2011

[4] Rehm G., Uszkoreit H. (eds), *META-NET White Paper Series: Europe's Languages in the Digital Age*. Springer, Heidelberg, 2012

[5] Rehm, G. (edt.) (2015). Strategic Agenda for the Multilingual Digital Single Market, Draft version 0.5, http://rigasummit2015.eu/sites/rigasummit2015.eu/files/Strategic-Agenda-for-Multilingual-DSM%20.pdf

[6] Rozis, R., Skadiņš, R., Vasiļjevs, A. (2016). Collecting Language Resources for the Latvian eGovernment Machine Translation Platform, Proceedings of LREC 2016, 1270-1276.

[7] Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In Proceedings of LREC 2012 (2012), 36-42.

[8] Krauwer, S. (2003). The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. Proceedings of SPECOM 2003, 8-15.