5.9 Language resources to make digital single market a reality\textsuperscript{128}

Why invest to promote language resources?

The META Language white papers series, published in autumn 2012 indicated that as many as 21 European languages (most of which are official EU languages) are facing “digital extinction” – meaning that relevant and up-to-date online content and online services are not available in those languages (or that the level of service in such “endangered” languages is significantly lower than English or other major languages).\textsuperscript{129} However, language technologies such as machine translation (MT) offer opportunities to overcome language barriers in the online domain. The problem is that for the smaller languages (especially for the 21 languages referred to above), the availability of MT solutions is limited, due to limited availability of language resources. This is because language resources are the indispensable “raw material” for producing high-quality MT systems. While EU-funded actions have been undertaken (see META: box below) to create a single access point to a European pool of language resources, this pool is not yet sufficiently populated by high-quality language resources for all EU languages. The possibilities for using EU research/innovation/infrastructure funds to support individual EU languages are limited, as such EU-funded projects typically focus on developing technology, infrastructures and solutions of generic and cross-border European value.

On the other hand, Member States and their regions are best placed to foster language resources for their respective languages. They also have an interest in promoting the status of their languages vis-à-vis other European languages. Very often, the under-resourced languages are also languages spoken in less developed regions.

A well-functioning and systematic use of ESIF to support language resources would contribute to the objectives of the DAE by removing language barriers from the online domain, helping to create a digital single market across the numerous linguistic borders. More concretely, it would complement the pan-European multilingual digital services foreseen in the Connecting Europe Facility, by ensuring coverage of all European languages, including the smallest ones, by complementing the “Automated translation” building block foreseen in the CEF proposal. It would also complement more ground-breaking work on improving machine translation quality, foreseen in the draft work programme 2014-15 for Challenge 4.3 of the H2020.

\begin{table}
\centering
\begin{tabular}{|l|}
\hline
\textbf{META}\textsuperscript{130} is a cluster of projects, aiming to build a Multilingual Europe Technology alliance (META-NET), and to create a pool of European language resources (META-SHARE).\textsuperscript{131} The project has also produced analyses (the Language White Papers) of the situation of EU languages in the digital and online context. META is complemented by regional projects coordinating the collection of language resources in their respective areas of coverage: \\
\hline
\textbullet \ META-NORD\textsuperscript{132} Baltic and Nordic language resources; \\
\textbullet \ CESAR\textsuperscript{133} Resources for Central and South-Eastern European languages (including Balkans). \\
\textbullet \ METANET4U\textsuperscript{134} Portuguese, Spanish, Romanian, English, Maltese language resources \\
\hline
\end{tabular}
\end{table}

Barriers & challenges

While EU projects from FP7 and CIP programmes have supported, to some extent, the initial organisation of language resource collection in Member States, the landscape of language resource coverage has numerous gaps. Typically, the biggest gaps appear for languages (like the East European languages) that have modest support from national budgets, while bigger language groups (EN, FR, ES) are better served and have more language resources, thus, also they are equipped with better automated multilingual solutions like MT.

The involvement of Member States and regions in programming, strategic planning and implementation of EU-funded actions in language technologies and language resources has so far been limited. In many Member States, it has been difficult or impossible to identify contact persons or desk officers that could be involved in EU projects, invited to events or informed of the results of projects.

How to act?

Regions wishing to invest in the language resources should consider the following steps:

1. **Analysis**: Member states and regions are in a key position for detecting, making aware and mobilising the relevant actors that hold linguistic material and key competences, or operate in local language services markets. The key stakeholder groups are:

\begin{itemize}
\item \textsuperscript{128} This area of activity can be an important element of a digital strategy. It might, however, not be eligible for funding through ESIF.
\item \textsuperscript{129} \url{http://www.meta-net.eu/whitepapers/press-release}.
\item \textsuperscript{130} \url{http://www.meta-net.eu}
\item \textsuperscript{131} \url{http://www.meta-share.eu}
\item \textsuperscript{132} \url{http://www.meta-nord.eu}
\item \textsuperscript{133} \url{http://cesar.nyutud.hu}
\item \textsuperscript{134} \url{http://www.metanet4u.eu}
\end{itemize}
• Public sector, such as national/regional ministries in charge of ICT, commerce, languages – but also those producing relevant language resources (e.g. translations)

• National language councils and similar bodies, tasked to preserve and develop national linguistic heritage and competences

• Universities and research centres engaged in research in linguistics, language technologies and language resources

• Companies that provide language services (e.g. translation, localisation) or language technologies (e.g. machine translation) for the local/national language(s) and/or for the local/national markets.

2. Governance/stakeholder involvement: It is important to ensure that Member States and regions are involved in and contribute to the planning and implementation of EU actions on language resources, as they have a direct interest in their respective languages. At minimum, Member States and regions should appoint desk officers or representatives to follow European actions and EU-funded projects that coordinate work on language resources and language technologies. Ideally, they would draw up national/regional programs and set up related governance structures to support their language in the digital context, and to seek collaboration with related EU actions.

Member States and regions should coordinate the participation of the stakeholders in their respective geographic areas to EU calls for proposals, calls for tender and to various awareness and coordination activities and events. The national centres in the META-NET can provide information and assistance and propose collaboration in on-going projects and other actions. Member States and regions can propose additional centres of excellence and contact points, especially in case of gaps in representation of a particular language or region.

3. Priority setting: Priority should be given to language resources that fill a gap in coverage (i.e. language(s) and/or topics that are not yet covered). Such gaps are identified in the META-NET Language White Papers. Another basis for prioritisation is to promote high-impact applications, for example, e-commerce sites or e-government systems that serve (or have a high potential to serve) a large number of customers from different language communities. Finally, priority should be given to language resources that serve important policy objectives either at the EU level (e.g. digital single market) and/or at the Member State/regional level.

4. Policy mix: Member States and regions should use the opportunities in policy making and legislation to promote their languages in the EU context. This is especially important for programmes that provide funding for multilingualism, language resources, or language technologies. Also, effective transposition and implementation of the Public Sector Information (PSI) directive will improve availability and re-use of language resources generated by the public sector in the Member States.

Member states and regions often hold linguistic content that has high potential re-use value as language resources. One example is the collections of documents and their translations (e.g. regulatory texts, informative publications, web site content in multiple languages...), as such “parallel corpora” can be used to develop machine translation systems. Another example includes bilingual or multilingual dictionaries, glossaries and terminologies. Such language resources often require conversion and processing to machine-readable format, and, sometimes, more time-consuming annotation, before they can be used to develop machine translation engines and other useful applications. This requires investment of human resources and/or funds.

5. Monitoring and evaluation: Performance can be measured at different levels in the value chain of language resources:

• Number of Member States/regions/language competence centres participating in language resource collection;

• Number of language resources collections, their typology (topical coverage, type of resource) and size (number of words, tokens, gigabytes, segments, records etc.);

• Number of applications (e.g. machine translation systems) developed with the language resources;

• Standardised indicators (e.g. BLEU metrics) measuring the quality of the applications developed with the language resources;

• Size of user base for the multilingual applications, developed with the language resources;

• Economic impact (e.g. revenue generated) by the applications developed with the language resources.

Further reading

http://s3platform.jrc.ec.europa.eu/language-resources