As part of META-NET, the CESAR project, coordinated by **Tamás Váradi**, aims to standardise language resources used across Europe to enable a better understanding of a wide range of communication activity, making knowledge sharing and transfer easier

Language Technology to tackle the Multilingual Challenge

As Thomas H. Davenport and Laurence Prusak point out in: *Working Knowledge: How Organizations Manage What They Know*: "People can't share knowledge if they don't speak a common language".

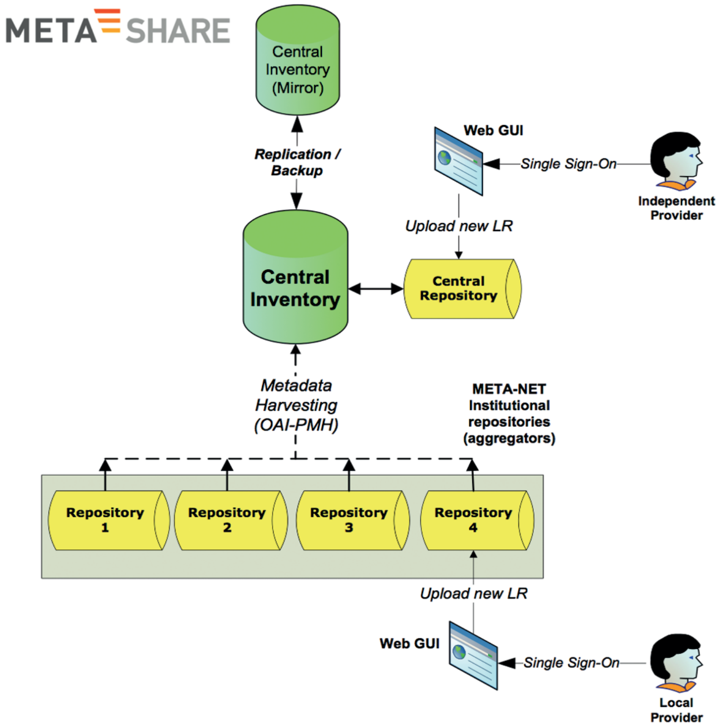
That is true across the world of course, but perhaps the observation's poignancy is felt most in the disparate societies of the European Union. From the West of Ireland across to the Eastern isles of Greece, from the southern tip of Spain to the north of Scotland, sooner or later there is going to be a need for us to communicate important and significant information with one another. Which is where the CEntral and South East EuropeAn Resources (CESAR) project comes in.

Led by its Hungarian co-ordinator, Tamás Váradi of the Research Institute for Linguistics at the Hungarian Academy of Sciences, and funded by the EC, through its ICT Policy Support Programme, CESAR is aimed at enhancing, upgrading, standardising, and cross-linking a wide variety of language resources and tools, as well as making them accessible, thereby contributing to an open linguistic infrastructure.

The project aims to make available a comprehensive set of language resources and tools covering the Bulgarian, Croatian, Hungarian, Polish, Serbian, and Slovak languages.

Resources include interoperable mono- and multilingual spoken and written databases, corpora, dictionaries and wordnets, as well as tools: tokenisers, lemmatisers, taggers and parsers.

CESAR itself is part of a larger EC project, called META-NET. One of the main objectives of META-NET is to produce a Strategic Research Agenda for Europe's Language Technology landscape. The agenda contains high-level recommendations, ideas for visionary LT-based applications and suggestions for joint actions to be presented to the EC and national as well as regional bodies.



Some four separate groups are involved in this vision-building process within the initiative. The main task of the three domain-specific Vision Groups (Translation and Localisation, Media and Information Services, Interactive Systems) is to produce ideas and concepts for visionary LT-based applications within their respective domains and to provide input for technology forecasts. The fourth group involved in the vision-building process is the META Technology Council which reports on the accumulated and condensed findings produced by the Vision Groups. The entire project is due to reach a conclusion by January 31, 2013.

According to a recent EC report 57 per cent of European Internet users purchase goods and services in other languages. Reading content in a foreign language is accepted by 55 per cent of users, while just 35 per cent use another language when writing emails or posting comments on the web. Moreover, just a few years ago, English was the lingua franca of the web. Today, that situation has drastically changed and the amount of online content in other languages, especially Asian and Arabic languages, has exploded. So Europe really needs to act now in order to prepare its 23 official languages and 60 spoken languages for the digital age.

Dr Váradi takes up the story: "META-NET is a network of excellence basically consists of one FP7 and three ICT-PSP projects that are individually funded.

"Three projects - three regional groupings - were initiated last year (2010) with six to 10 partners at most covering five to six languages each, so that there should be one or two partners per language at most and that these groupings should be either regional like ours, or perhaps linguistic, so that all the Romance languages would form a consortium, but the idea was to give support to the notion – and this is one of the main missions of all three of these projects, so I am now placing CESAR in the larger picture – that we should build META-NET.”

So did the project have antecedents? Surely there had been a number of basic attempts to standardise communication throughout the EC in the past?

“Yes, the idea is that there were, for a number of years, separate isolated attempts at various research centres to develop what we call language resources – textual databases, a corpus in fact, which is a collection of texts, uniformly annotated and with additional linguistic analysis or information added.

“These language resources are valuable assets but in order to exploit their true potential for European-wide research and development they need to be brought to a standard format, enhanced, aligned and made available in a distribution facility.

“One of the overall goals of META-NET, through what is called META-SHARE, is to build a distributed system whereby these resources are made available for download for industry or for research and development purposes. Language technology is an all-encompassing field because language is an innately and inherently human facility.”

Language is a universal faculty of humans but languages also vastly differ, as we all know when we try to speak another language. So how are individual differences between languages taken care of?

“In order to develop, let’s say a French to Lithuanian machine translation, you need to have a huge parallel corpus of translated French text translated into Lithuanian and vice versa,” she continues. “So these language resources are the data that provides for taking care of the individual differences between languages. The idea is that the tool can become language-independent at this abstract level but in order to derive with the tool some specific application, like the French to Lithuanian translation system, you need to feed it a huge amount of specific data, the more the better.

“This is how, on a very general level, the differences between individual languages are being taken care of by supplying the relevant specific, carefully compiled and analysed language data, relating to particular languages.

“ =A lot of emphasis is on the separation between the algorithm on the one hand and the data driving this in particular languages.”

The ambitious two-year project only began in February this year and is currently at a fairly preliminary stage with an initial list of resources at the preparation stage.

“The mission of our participating nine centres is not only to prepare and adjust our own resources that we ourselves provided as language-resource providers but our duty is to act as a catalytic force and reach out and motivate the rest of the language technology centre for the particular language and to approach partner institutes and also get them motivated and involve their resources into this META-SHARE system, mainly to make them accessible and available.

“Now there is a standardised way to make such information available and it is of the utmost importance when you build such a network of distribution centres, through which you can have access to these resources, that these metadata should be uniform so that you can query it, you can locate it and then you can download it or use it online.

“This is the idea, so you can build applications and so the main mission of these participating META-NET projects, including CESAR is to locate these resources and tools, to engage in collaboration with partners of the relevant language and the technical work relates to preparing the resources and tools so that it can become part of this pool of resources and tools distributed through META-SHARE.

“Our other mission is not just to engage with other partners – providers of resources and tools – but also to raise awareness of the potential in language technology with the media, with the policymakers, because although this is a two-year project, we certainly plan for a long-term period, although the EC might not necessarily finance us certainly not forever.

“Therefore, the other very important objective is to find interest and to raise motivation of national funders to invest in supporting such an effort and supporting language technology in general,

“Our next big event is Metaforum 2011, a major conference and show, involving not just the research and development sectors, but also industry and we would also like to generate interest from politicians, because this is the big opportunity for all stakeholders in language technology, information providers and users in industry to get together and exchange very important and pertinent ideas.”

At a glance

Project Information

Project Title:
CESAR Central and South Eastern European Language Resources

Project Objective:
The CESAR project, as a part of META-NET intends to enhance, upgrade, standardise and cross-link a wide variety of language resources and tools and making them available to the public, contributing to an open linguistic infrastructure.

Project Duration and Timing:
1 February, 2011 – 31 January, 2013

Project Funding:
CT-PSP Programme,
Total project cost: 4160002 €, 50% funding by EC

Project Coordinator:
Research Institute for Linguistics, Hungarian Academy of Sciences

Dr. Tamás Váradi

In 1997 he founded the Department of Corpus Linguistics (later called Department of Language Technology), whose initial mission was the creation of the Hungarian National Corpus. Under Dr. Váradi’s supervision the department has participated in a number of major national and EU funded research projects in the area of machine translation, information extraction, language resources building. Dr. Váradi has gained recognition in the area of language resources and language technology across Europe. He regularly serves on the program committees of several conferences including EACL, TSD, LREC, and Digital Humanities. He acted as local host to EACL’. He was appointed President of the TELRI (Trans-Language Resources Infrastructure) Association in 2003. Recently, he was one of the founding partners of the major research infrastructure project CLARIN (www.clarin.eu) and he is a member of the Executive Board. He also acts as the presidents of the Hungarian Technology Platform for Language and Speech Technology.

member of
META-NET

Contact

Main contact name:
Tamás Váradi
Tel: 0036 321 4830 ext. 126
Email: varadi@nytud.hu