Tamás Váradi

The relevance of language technology infrastructures: national and European initiatives

Abstract

In this short paper I intend to show the increasing importance of language technology as infrastructure that can support research and development and various ICT applications. I will present two large scale European projects (CLARIN and CESAR) and two examples from the Hungarian scene (The Language and Speech Technology Platform and the National Register of Research Infrastructure). Finally, I will discuss the relevance of these initiatives for EFNIL.

E rövid dolgozat célja, hogy bemutassa, hogy a nyelvtechnológia mint infrastuktúra egyre fontosabbá válik a kutatás-fejlesztés és különböző információ-technológiai alkalmazások támogatásában. Mindezt két nagyszabású európai (CLARIN és CESAR) valamint két magyar projekt (Magyar nyelv- és beszédtechnológiai platform valamint a Nemzeti társadalomtudományi hivatkozás adatbázis) ismertetésével illusztrálom. A dolgozat végén röviden utalok ezen munkálatok relevanciájára az EFNIL számára.

1. The mission of language technology

It requires little reflection to realise that in our age communication is increasingly digital. Whether we already live in information societies is a moot point. Almost exclusively, we already use digital technology to talk and write to each other through electronic devices (mobile phones, computers, mobile various communication devices) in our personal lives. They all generate a huge amount of texts (to consider, for simplicity, just the written medium). On a larger scale, we find that in an increasingly globalised world, digital information is generated at a rate that threatens with information explosion. It becomes impossible to keep pace with the amount of information that is created in the media, science, economy, wherever we look, in fact.

Despite the prominent role of multimedia, human communication is and will always be based on language, a facility that is widely held to be an innate characteristic of humans. Language is so intricately involved in thinking and the whole human existence that it is inconceivable that human communication will be conducted in any other medium.

This situation presents an enormous challenge to language technology, a multidisciplinary field comprising of computer science, computational linguistics, artificial intelligence, psychology etc. If we use machines to communicate with each other, we must enable these machines to process language with the same ease and intelligence that humans do. In other words, we must equip them with linguistic knowledge and intelligence that, ideally, approximates the linguistic competence of humans. In a sense, this is a futuristic goal converging with the vision of artificial intelligence. Some people may not even like positing such goals, contemplating with abhorrence the idea of thinking machines. We need not be too much concerned about the philosophical implications as it is doubtful if, in principle, this aim can be realised at all. On the other hand, the pressing global need for facilitating human communication via machines is undeniable and is already an every-

day experience. Making machines more adept at processing language helps us to communicate with machines. In other words, it increasingly frees us from the constraints imposed on us by limitations of the hardware and the operations of the machines. But language technology not only serves the purposes of human-machine communication since we use machines nowadays for human to human communication, therefore, a major part of the mission of language technology is to serve human communication in general.

Language technology may not be a familiar field, yet its results are already with us. Spell checkers, scanners that recognise texts (optical character recognition systems), internet search engines and particularly machine translation, these are all examples of what language technology can do to facilitate human communication. None of these technologies is perfect, yet all of them already serve their purpose and, indeed, we'd immediately feel their absence if we did not have recourse to them.

2. Language technology as infrastructure

Language technology should aid us to create, translate and summarize texts. (I am using text as a cover term to refer to language output whether written or spoken.) A very important requirement in this age of information explosion is to find relevant information in free text and organize it into useful knowledge. With respect to speech, it would be extremely useful if machines understood what we say, at least in some basic sense of the word and if they responded to it in an intelligent way and if they were able to speak to us in a natural manner.

It is important to realise that all the above general requirements are domain independent tasks. This leads us to suggest that the provision of all these facilities should be considered part and parcel of the services that digital communication technology should provide. In other words, language technology should be regarded as a part of the *infrastructure* that we use in modern information communication technology (ICT). If this proposal needs justification, let us just consider what good it is to bring broadband internet access to the remotest corners if the language barrier does not make accessible the content of what becomes available down the lines.

This is the concept of language technology as infrastructure at the most basic layer of ICT. It is a long-term vision but, as we saw earlier, elements of this infrastructure are increasingly becoming reality.

There is another sense in which language technology is already recognised as infrastructure and, indeed, is being developed under various European and national initiatives that will be described in the next two sections. This is at one remove from being deployed in front-end applications. One such infrastructure (CLARIN) serves the purposes of scientific research, and within it, scholarly research in the humanities, in particular. Another major on-going project (CESAR) intends to foster multilingual Europe through the provision of language resources and tools in a standard format widely distributed in a dedicated network of exchange facilities.

3. CLARIN

The CLARIN infrastructure was called to life by ESFRI (European Strategic Forum for Research Infrastructure), a European political initiative that was set up in 2002 following a decision by the Council of Ministers "to support a coherent and strategy-led approach to policy-making on research infrastructures in Europe and to facilitate multilateral initiatives leading to a better use and development of research infrastructures".[1] The newly formed body proceeded to compile a roadmap of European Research Infrastructures out of infrastructure proposals submitted in response to a call and which was judged by independent peer review. CLARIN was born as a result of the merger of three language technology proposals and became one of the six proposals selected in the social sciences and humanities category.[2] The initiatives in the ESFRI Roadmap were invited in a closed call to submit project proposals to DG Research and Innovation. As a result, the CLARIN project was launched in 2008 with the coordination of Steven Krauwer of Utrecht University involving 35 partners from 25 countries.

The CLARIN project (www.clarin.eu) has the ambitious long-term mission to develop a distributed infrastructure that would serve ultimately as a virtual research environment in which the users could benefit from language resources and tools as well as advice on how to apply them to the research questions at hand. CLARIN intends to focus primarily on scholars in the humanities and social sciences as they were judged to require special attention for the following reasons: their work typically involves texts, which are increasingly available in vast quantities in electronic format. Research in the humanities is carried out as individual efforts by scholars who are relatively less familiar with the benefits of language technology. In addition to the perceived needs and requirements of the target audience, the CLARIN infrastructure is also motivated by the fragmented nature of the language technology sector. In particular, it was noted that there is a huge number of language resources and tools that were developed as isolated efforts, with little regard to standardised formats and the additional benefits that come from interoperability, the possibility that any particular tool can operate with a variety of resources. The planned infrastructure would locate these isolated centres and would make their tools and language resources available in a unified framework for the benefit of the humanities scholar.

CLARIN as an EU-funded project is only the beginning, the preparatory phase of an open-ended enterprise. The EC provided only the seed money, formally only to work out the legal, organisational and governance structure of the future infrastructure. The preparatory phase is to enter the period of the construction of the infrastructure, designed to last for five years and funded exclusively by the member states. Since the launch of the ESFRI infrastructure projects the EC has created a special European legal entity, called ERIC, European Research Infrastructure Consortium. CLARIN is currently transforming itself into CLARIN ERIC and the construction of the infrastructure is about to begin early 2012.

---

1 Cf. http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri-background.

2 http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri-roadmap&section=roadmap-2006.

While the national support in the form of governmental commitment proved difficult to secure for the majority of the CLARIN consortium members (CLARIN ERIC is likely to begin its operation with at most 14 members), CLARIN itself has attracted enormous community support. The number of research centres that aligned themselves with the aims of CLARIN is now above 200 and still growing. The register of mono and multilingual corpora, lexica as well as processing tools is truly impressive. Their list is available with faceted browsing at http://clarin.eu/vlo.

The activity within CLARIN in the preparatory phase did not confine itself to the formal aims of the original call for proposals. Indeed, the CLARIN project proceeded to develop a prototype of the planned infrastructure. It tackled this complex task in several dimensions. There was the task of building the technical backbone of the infrastructure. When fully completed in the construction phase, the CLARIN infrastructure will allow the individual humanist scholars to access a large range of resources residing at different centres across Europe and process them using tools supplied by other centres, all this achieved in a seamless operation from the comfort of their laptop. A schematic view of the intended CLARIN infrastructure is displayed in Figure 1.

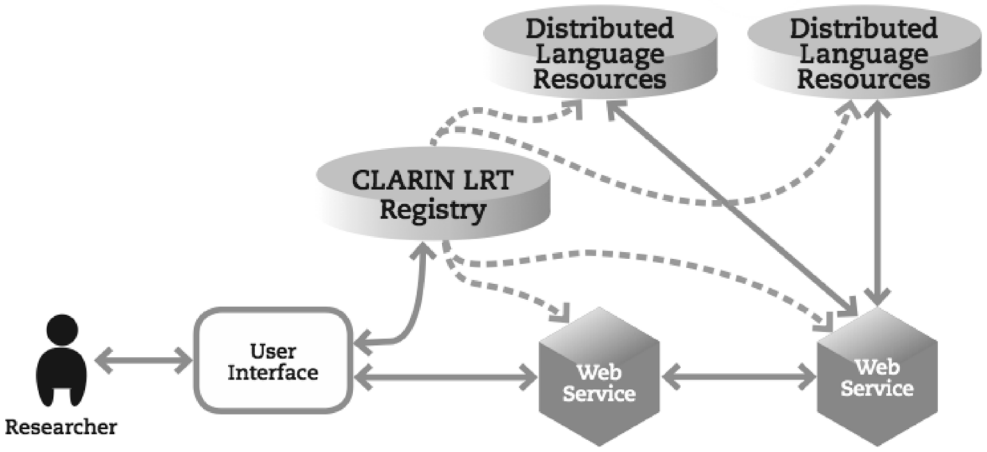

Fig. 1: Overview of the planned CLARIN infrastructure

The second pillar of the infrastructure relates to the content that will be made available down the electronic pipelines. This raises the issue of the language policy that CLARIN has adopted. Being an all-European research infrastructure, CLARIN is committed to supporting all the official national languages small and large, with particular attention to the former. In addition, CLARIN has to extend the linguistic horizon both in space and time if it wants to cater to the needs of humanist scholars. In other words, it has to provide support, on the one hand, for historical stages of modern European languages and, on the other hand, for languages that are not native to Europe but have a great tradition of being studied in leading European centres.

The third pillar can be aptly called the knowledge infrastructure. Right from the start, it was realised that the target audience needs advice and guidance in applying language technology to solve the particular research questions they are concerned with. An important activity, then, was to reach out to the community of scholars, to understand their research concerns, methodology and to find out about potential bottleneck in their uptake of the technology developed within CLARIN. The main instruments CLARIN decided to

adopt in pursuit of the above aims was to create a large-scale survey of relevant organisations, projects and conferences, to liaise with professional humanities organisations and to engage in actual collaboration with individual projects. To the latter end, CLARIN selected a handful of projects through an open call and supported them by advising on how to apply language technology to realise their objectives. Making a large-scale impact on the target community proved an extremely difficult task, yet working with selected groups of researchers turned out to be a mutually rewarding task.

4. CESAR and META-NET

The CESAR (Central and South-Eastern European Language Resources) infrastructure (www.meta-net.eu/projects/cesar/) is a two-year ICT-PSP project consisting of nine partners from six countries (Poland, Slovakia, Hungary, Croatia, Serbia and Bulgaria) coordinated by the Research Institute for Linguistics, Hungarian Academy of Sciences that started its work in February 2011. One of the main objectives of the project is to make available language resources and tools that exist within the respective language technology community properly documented, equipped with a rich amount of metadata and cross-linked, where possible, to ensure they are interoperable. The resources and tools will be contributions to an open language resource infrastructure. The CESAR project is part of a larger initiative called META-NET (META standing for Multilingual Europe Technical Alliance) that is now a growing alliance that aims to reach all stakeholders interested in fostering multilingual Europe through modern language technology. META-NET currently includes 47 members from 31 European countries.

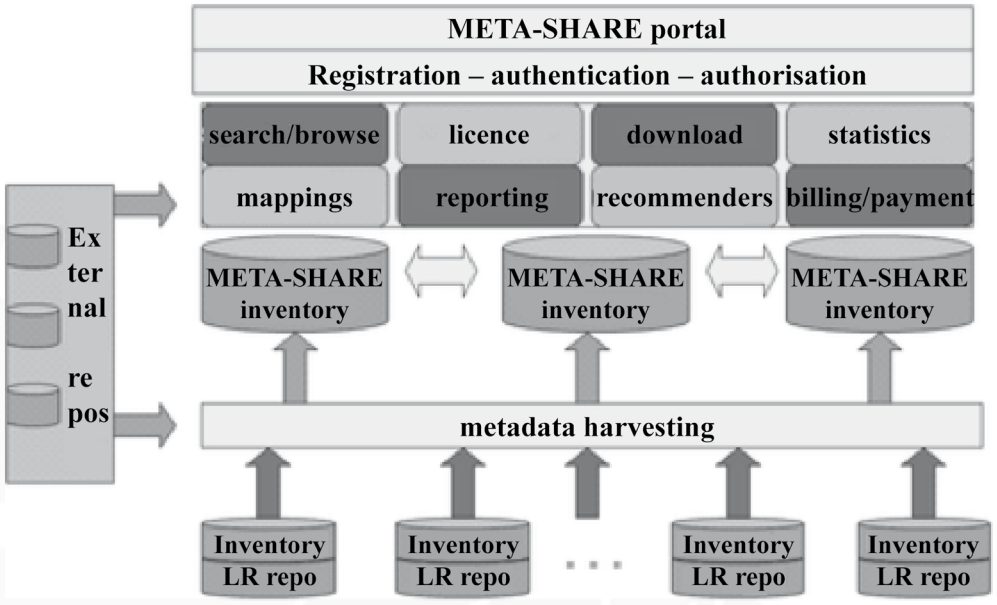

Fig. 2: Schematic view of META-SHARE

META-NET is not a research infrastructure, it rather aspires to build a large-scale alliance of technology partners, industry, policy makers and corporate and individual users. While its overall ambitions are rather general, it tends to foster the cause of multilinguality by providing support for the development of online web services. The paradigm application is widely considered to be statistical machine translation (SMT). Machine

translation is the pinnacle of what language technology can offer and requires complex technology as well as an enormous amount of data. The pooling of language resources and tools is one of the central activities in META-NET and they will be made available in a distributed network of repositories organised in the META-SHARE system.

Following the rules of the call for proposal, the CESAR project involves one or two partners per language. However, their mandate is to act as a catalytic force and mobilize all stakeholders of the language technology scene of the respective language including not just research and development centres but industrial partners, potential users and policy makers and the media. Indeed, one of the measures of success of their activities as contributors to META-SHARE will be the extent to which they will be able to increase their portfolio of resources with those that come from partners. In this context, EFNIL institutions as owners and providers of valuable resources of the national languages across all Europe are important strategic partners for META-NET.

Figure 3 shows (at the top) the strategic documents META-NET plans to produce as well as (at the bottom) the process of planned consultation and communication that is leading to and follows the creation of these documents. The presentation given by Hans Uszkoreit, coordinator of META-NET at this conference is part of these efforts.

Fig. 3: Timeline of the META-NET agenda

5. Hungarian Language and Speech Technology Platform

The Europe-wide project META-NET had a close parallel on the national scene in the form of the creation of technology platforms. The technology platform as an industry-led instrument to strengthen the European Research Area was called to life by the European Commission in 2003.[3] The European initiative was followed up in the member states and as a result, the National Office for Research and Technology (currently renamed to National Innovation Office) issued the first call for proposals to form national technology platforms. The objectives of the projects were to be the following:

– Unite and mobilize all major technology partners in a given field;
– Develop a Strategic Research Agenda;
– Work out Implementation Plan based on SRA;
– Raise awareness of the field (public, media, policy makers);
– Reach out to major stakeholders in the sector.

[3] http://cordis.europa.eu/technology-platforms/about_en.html.

The nature or the size of the sector was not defined and the first ten winning projects included platforms of hugely different sizes. The rationale for the national technology platforms was that it should enable stakeholders in particular research and development fields to organize themselves in a bottom-up way and to define their own strategic vision for themselves as well as to compile an implementation plan. These documents would inform policy makers who would base national strategic research and development plans on the SRA's of the national platforms.

The Hungarian Language and Speech Technology Platform (www.hlt-platform.hu) was founded by four academic and four industrial partners. Most of them had been collaborating in various projects in the past few years so it was a tested and tried consortium led by the Research Institute for Linguistics. The technology platform was welcomed as an excellent opportunity to engage in ancillary activities that go beyond the scope of ordinary R&D projects such as strategic planning and large scale PR activities. The project staged three high profile events in the form of public conferences, the first one to introduce the Platform and the achievements of Hungarian language technology, the second conference focused on the Strategic Research Agenda and the third major publicity event introduced the Implementation Plan. Each event included a demo session where the latest language technology developments were showcased.

The two year platform ended with all the goals of the project successfully completed. The visibility of the mission and potential of language technology was greatly enhanced as evidenced by the fact that the hugely popular Hungarian open university television series *Mindentudás Egyeteme* included language technology as one of the first subjects covered in its recently opened second season.[4] The members of the Platform more than doubled and the majority of the new members came from small and medium size enterprises. The major achievements of the project included the Strategic Research Agenda[5] and the Implementation Plan,[6] which were compiled and submitted to public debate on the website and the two conferences.

6. Bibliographic Reference Database for the Humanities

The idea for this project arose when the Initial List of the European Reference Index for the Humanities (ERIH)[7] was published in 2009. The purpose of ERIH is to increase visibility of European Humanities research and introduce some solid measuring criteria of evaluating research output. It was compiled through a Europe-wide community effort coordinated by ESF.

The Bibliographic Reference Database for the Humanities project was inspired by the general objectives of the ERIH. Research results in the Humanities suffered from the same lack of recognised standards of quality and the resultant low prestige with respect to natural sciences, for example. In addition, in the absence of a central reference database,

[4] http://mindentudas.hu/elodasok-cikkek/item/2520-sz%C3%B3b%C3%B3l-%C3%A9rt?-%E2%80%93-emberg%C3%A9p-nyelvtechnol%C3%B3gia.html.

[5] http://www.hlt-platform.hu/skt.

[6] http://www.hlt-platform.hu/sites/default/files/MT_vegleges.pdf.

[7] http://www.esf.org/research-areas/humanities/erih-european-reference-index-for-the-humanities.html.

humanities researchers are forced to compile the list of references to their publications, which they are often ill-equipped to carry out and most of them consider it an unnecessary burden at best. The projected Reference Database aims to cover the comprehensive list of Humanities journals published in Hungary. The scope of the database had to be carefully defined both in terms of geographical and chronological dimensions. For practical constraints, inclusion of references to Hungarian journals published abroad could not be considered. Coverage of journals would start with recent numbers and proceed in reverse chronological order. As the Reference Database is expected to serve the very practical scientometrical requirements of the present-day generation of Humanities research, we do not expect to go back in time longer than the stretch covering living authors.

The Reference Database would serve as a metric not only for authors but at the same time for journals themselves and is widely welcomed by librarians, publishers, administrators and officials at universities as well as the Hungarian Academy of Sciences. The Research Institute for Linguistics has decided to launch this project because, although the work is complex and involves the deployment of robust hardware and software technologies, it crucially depends on language technology. The challenge is to parse the citations that appear either appended to the articles or at the bottom of the page and convert them into structured information. While the citations may have originated a bibliographical database, they are published in more or less free form as text. Although journals typically publish style sheets containing instructions for the format of bibliographical entries and indeed there are a great number of standard citation formats widely published and used in a number of journals, our initial findings indicate that, unfortunately, Hungarian journals in the humanities are very slack in enforcing a standard form even within the same journal.

Bibliographic references seemingly represent a fairly closed format and humans are very good at understanding them at a glance. Nevertheless, processing them with computers presents technological challenges. Even if some standard format is followed (which, unfortunately, cannot be taken for granted) the title field of the citation can hardly be processed adequately without a measure of understanding it. This, however, typically goes beyond current technology, therefore lack of deep processing of the title must be compensated for with some heuristics. It must be accepted, nevertheless, that automatic processing will have to be complemented with manual effort, the crucial question is rather the extent to which the work will have to rely on manual work.

7. Conclusions

In the above sections we described four language technology projects that vary in scope and domain but all provide valuable infrastructure. In this concluding section, we consider the relevance and implications of these projects for EFNIL.

First of all, the relevance of EFNIL can be twofold, depending on the two kinds of members that make up EFNIL's strength. EFNIL is unique in that it unites national language institutes as well as representatives of organisations dealing language policy and language planning.

National language institutes are typically the centres where the major language resources such as dictionaries, corpora and other collection of valuable linguistic datasets are produced. In fact, often their fundamental mission centres on the creation and publication of these resources. Therefore, they should be inherently interested in seeing that their resources are actively used among the widest possible audience. In this increasingly digital age, this goal can only be ensured by dissemination methods using modern technology. On the other hand, the technological know-how and facilities are often not available at EFNIL institutions – rightly so, we might add, as such activities fall outside their core agenda. It is all the more important and opportune that EFNIL members as providers of invaluable and often unique language resources of the respective language should join infrastructures such as CLARIN and META-NET in order to use their distribution and data curation services. Fortunately, quite a number of EFNIL institutions are already participating in one or both of these infrastructure projects.

Policy makers responsible for language policy within particular EFNIL member states can be most efficient partners to EFNIL institutes in their effort to join these infrastructures. CLARIN is no longer a project but will resume its operations as CLARIN ERIC, which entirely depends on national support at governmental level. Clearly, EFNIL members representing relevant governmental organisations having a clear understanding of the goals and importance of language technology infrastructure can further the dissemination objectives of partner EFNIL institutions.

The two national projects also bear some relevance to EFNIL members in that they can be implemented in other member states. The work on the Reference Database is also eminently suitable to scaling up, preferably in a coordinated way as a pan-European effort culminating in a European Reference Database for the Humanities.

In conclusion, it is hoped that this brief overview has shown how language technology infrastructure can be useful in furthering the general objectives of EFNIL and why it is therefore important for individual EFNIL institutes and organisations on the one hand and EFNIL as an organisation on the other to cooperate with current infrastructure initiatives on the national and European level.