CLARIN

A European Research Infrastructure

Newsletter

Number 3, 2008, October

Language Infrastructures: what happens outside EU?

Nicoletta Calzolari

ILC-CNR, Pisa, Italy

The setup of CLARIN in Europe was the result of a long series of initiatives and attempts from many of us, starting already at the beginning on the 6th Framework Programme. That time is finally ripe for such an infrastructure is shown also by other initiatives outside Europe that share objectives and ideas with CLARIN. I mention here just a few.

Probably the most similar is a 5-year program just finished in Japan: the 21st Century COE (Center of Excellence) Program "Framework for Systematization and Application of Large-scale Knowledge Resources", led by Sadaoki Furui at Tokyo Institute of Technology[1]. It aimed at systematising and relating a variety of multimedia information to make use of them as 'knowledge': this is one of the key issues in the 21st century. I was member of the International Board and could observe the breadth of the areas covered to construct, integrate and use large-scale knowledge resources (from spontaneous speech, written language, to materials for e-learning and multimedia teaching, classical literature and historical documents) in many domains of research for Human and Social Sciences. I also witnessed one of the most difficult problems in such interdisciplinary research combining humanities and technology, i.e. communication among researchers belonging to different communities and with very diverse backgrounds. I think this is a problem that CLARIN should expect and to which it must dedicate attention now already.

Another ongoing Japanese project with similarities with CLARIN is Language Grid, led by Toru Ishida at the National Institute of Information and Communications Technology (NICT)[2] and Kyoto University[3], with partners also in Europe (DFKI, ELDA, ILC-CNR). The Language Grid is an infrastructure built on top of the Internet to allow not only professionals but also end users to conquer the language barriers. The project includes language resource (LR) and computation resource providers, as well as language service users, and is based on Web service technologies that enable users to freely combine software distributed via the Internet. Semantic Web technologies enable the collaboration among LRs and language processing functions for intercultural activities, to improve the accessibility and usability of existing language services and to easily develop new language services by combining existing ones. It also offers an infrastructure where stakeholders can provide and/or use LRs by mutual consent, with understanding and resolution of the intellectual property issues. The main goal is to allow a better understanding of Internet content written in different languages and by people from different countries.

Participants of Unified Linguistics Anotation Workshop (http://verbs.colorado.edu/ula2008/)

I add to this picture an US-funded effort, the NSF CISE-CRI (Computer and Information Science and Engineering-Computing Research Infrastructure) "Towards a Unified Linguistic Annotation", led by James Pustejovsky at Brandeis, with Martha Palmer, Adam Meyers, Mitch Marcus, Aravind Joshi and Jan Wiebe. This project, developing a Unified Linguistic Annotation (ULA) that integrates in one framework different layers of annotation (e.g., semantics, discourse, temporal, opinions) and several existing resources, including PropBank, NomBank, TimeBank, Penn Discourse Treebank, and coreference and opinion annotations, aims at providing a large corpus with balanced and annotated data. The project also aims at achieving an international consensus on a meta-specification framework allowing individual annotations to cohabit with each other, as well as a language-independent methodology and widely accessible tools and guidelines. The activity enhances infrastructure for research and education by providing a resource that could lead to major advances in robust, broad coverage semantic processing.

The set of these initiatives, sharing partially similar perspectives and highlighting the value and the need of new language infrastructures is, on one side, a sign of the timeliness of the CLARIN effort, and, on the other, invites all of us to a more global collaboration. As a last remark, I add that efforts toward the cooperation of these and similar initiatives all over the world will also be one of the aims and tasks of two new projects, the European e-Content-plus FLaReNet (Fostering Language Resources) Thematic Network, led by me, and the American NSF INTEROP project, just started, led by Nancy Ide and James Pustejovsky. Worldwide collaboration on these infrastructural issues is an essential step towards better exploitation of all the resources and technologies we develop and therefore towards higher impact of our field in the society.

[1] http://www.coe21-lkr.titech.ac.jp/

[2] http://langrid.nict.go.jp/en/

[3] http://www.langrid.org/association/indexe.html