Organization Models for Research Infrastructure and existing infrastructures

Bente Maegaard

Center for Sprogteknologi, University of Copenhagen
Njalsgade 80, DK-2300 Copenhagen
bente@cst.dk

Abstract

Since the 1990es, various organizations have been taking care of the distribution of language resources for research and commercial applications. Technical developments recently have opened new possibilities; how do we organize ourselves in the future? do we need new organizations, or modifications to existing ones?

1. Background

Ever since computers were born, there has been a need to collect and analyze language resources. Most of the very first applications of computers were corpus investigations. Some early investigations worked on what was felt then to be pretty large corpora, e.g. the Brown Corpus of 1 million words (Kucera & Francis 1967), and Kierkegaard's works of 2 million words (McKinnon 1965).

At that time, dedicated researchers typed in text themselves, or rather, raised money to have text typed. When the corpus was available, it was often accessible to only one or a few researchers for their own work. But even if they wanted to share with others, the technical means, the infrastructure and the copyright problems were too important obstacles. The users were not only linguists, but also historians, philologists etc.

1.1. The Index Thomisticus

As an example of a very early resource project, which was not necessarily linguistic, let us consider the Index Thomisticus. In 1946 Father Busa planned the Index Thomisticus, as a tool for performing text searches within the massive corpus of Aquinas's works. In 1949 he met with Thomas Watson Sr., the founder of IBM, and was able to persuade him to sponsor the Index Thomisticus. The project lasted about 30 years, and eventually produced in the seventies the 56 printed volumes of the Index Thomisticus. In 1989 a CD-ROM version followed, and a DVD version is underway. In addition, in 2005 a web-based version made its debut, sponsored by the Fundación Tomás de Aquino and CAEL. This is an example of a huge amount of work which was sponsored from the very beginning, and which was shared first through printed books and then through CD distribution, - when the technical possibilities were available.

1.2. Infrastructure in the previous century – what is the problem?

The infrastructure problems mentioned above include e.g. the fact that as soon as a resource is to be distributed, it needs to be in good shape: a clean version has to be made, and it has to be accompanied by documentation in some widely known language etc. Still today, the reason that many resources are not distributed is that it takes some energy to prepare them for distribution, and this is work that has to be done by those who produced the resource and therefore know it. However, the efforts that are needed to prepare a resource for sharing are much smaller than the effort to build the resource again, so researchers should be encouraged to make this last investment in their resource.

For commercial applications, the problems are different. If a company has built a resource they do not necessarily want to share it with others, as the resource may provide a competitive advantage. Below, we are focusing on research use of resources.

For research, shared resources provide benefits that 'private' resources do not, apart from the fact that more researchers can use the same resources. Shared resources also permit replication of published results, support fair comparison of alternative algorithms or systems, and permit the research community to benefit from corrections and additions provided by individual users.

2. Existing infrastructures

As an answer to this arising understanding of the possibilities in shared resources, two organizations were established in the 1990es, LDC and ELRA. We first present LDC shortly, and then go into more details with ELRA.

2.1. LDC

The Linguistic Data Consortium (LDC) was founded in 1992 to provide a new mechanism for large-scale development and widespread sharing of resources for research in linguistic technologies. Based at the University of Pennsylvania, the LDC is a broadly-based consortium that now includes more than 100 companies, universities, and government agencies

The Linguistic Data Consortium is an open consortium of universities, companies and government research laboratories. It creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes. The LDC was founded in 1992 with a grant from the Advanced Research Projects Agency (ARPA), and is partly supported by grant IRI-9528587 from the Information and Intelligent Systems division of the National Science Foundation.

2.2. ELRA

In Europe, the European Language Resources Association (ELRA) was founded in 1995.

Antonio Zampolli was the main driving force behind the creation of ELRA. The starting point was the realisation that the development of language technologies

was crucially dependent on the capability of processing large quantities of 'real' texts and on the availability of large-scale lexicons. This gave rise to the so-called 'reusability' notion which was at the basis of many initiatives for establishing standards and best practices.

This trend arose also from the increasing interest of national and international authorities in the potential of the so-called 'language industry'. The path went through a wide range of language resources (LR) projects, most of them financed by the European Commission (EC), both projects that aimed at developing LRs, and projects that were of a more political and coordinating nature. Within the EC Language Engineering (LE) program there was a very fruitful combination of LR, language technology and application projects, recognising the natural links among these aspects and need for them to proceed in parallel, in synergy, and in a coherent way.

Zampolli clearly delineated the major strategic lines of activity:

- elaboration of consensual standards,
- creation of the necessary LRs,
- distribution and sharing of LRs,
- creation of synergies among national projects, European and international projects, industrial initiatives.

In order to carry out such a strategic analysis, A. Zampolli, together with a large number of key players in the European language technology field proposed to the European Commission to launch a project called RELATOR - A European Network of Repositories for Linguistic Resources (1993-95). The project aimed at defining a broad organisational framework for the creation of the LRs, for both written and spoken language technology, which are necessary for the development of an adequate language technology and industry in Europe. It also aimed at determining the feasibility of creating a coordinated European network of partners that would perform the function of storing, disseminating and maintaining such resources.

The major outcome of RELATOR was the creation of ELRA as well as the initiation of several Language Resource production projects (e.g. SpeechDat family, PAROLE/SIMPLE, POINTER, etc.). The RELATOR project presented final recommendations for establishing a collaborative infrastructure that would act as a collection, verification, management and dissemination centre, built on the foundation provided by existing European structures and organisations. RELATOR proposed the foundation of a European Association for Language Resources, which was registered in Luxemburg (**ELRA - European Language Resources Association**) in February 1995. ELRA was established as an independent, not-for-profit, membership-driven association. ELRA was supported by the European Commission through project funding in the first years, but has been self-supporting since 1998.

ELRA's initial mission was to set up a centralised not-for-profit organisation for the collection, distribution, and validation of speech, text, terminology resources and tools. In order to play this role of a central repository, ELRA had to address issues of various nature such as technical and logistic problems, commercial issues (prices, fees, royalties), legal issues (licensing, Intellectual Property Rights), and information dissemination. ELDA (Evaluation and Language Resources Distribution Agency) was established as the operational unit of ELRA.

The mission of ELRA is to promote language resources and evaluation for the Human Language Technology (HLT) sector in all their forms and all their uses, in a European context. Consequently the goals are: to coordinate and carry out identification, production, validation, distribution, standardisation of LRs, as well as support for evaluation of systems, products, tools, etc..- related to language resources.

3. New challenges

Lately, the field of computational linguistics has seen a number of new developments.

New types of resources are needed for language technology research and applications, e.g. multimodal resources. At the same time other fields of application than computational linguistics and language technology are seeing the advantages of computational access to resources, - history, philosophy, music, literature etc. This means that other types of resources have to be made available. Knowledge of the field is necessary to make the right resources available in the right form.

Another development is the presence of the Internet with masses of data. The Internet has become a major source of data for many researchers, and this will certainly continue. However, even if for some applications this type of data is acceptable, the data do not come with quality assurance, and e.g. free lexica on the Internet are not of the quality needed for most applications. Also, there are copyright issues to be solved when data are taken from the Internet. It can be assumed that quality of what is available on the Internet will grow as it has done until now, but to solve the copyright issues a political effort is necessary.

The Internet and GRID technologies also provide new possibilities for distribution of data. ELRA has e.g. almost exclusively been using CD, because many resources are too large to be downloaded through the web, - but new technology will change this.

4. Organizational models

The existing structures, LDC and ELRA, are different:

LDC is a consortium that an organization (university, company) may join by paying a subscription fee. The organization then receives all resources built during the year of subscription.

ELRA is a member-driven association. Members pay the membership fee, and may purchase resources at reduced prices. A good deal of the research resources are extremely cheap, but ELRA also provides resources for industry which are more expensive.

The difference between LDC and ELRA are 1) the consortium vs. association, 2) the fee structure.

LDC and ELRA have more similarities than differences: they both provide a legal framework for copyright and licensing issues, they both maintain a catalogue of available resources, they both support the development of and adherence to standards, they both ensure some kind of quality in their resources. Both entities also identify new interesting resources for their customers.

ELRA has set up formal procedures for validation of resources and made the validation manuals public. ELRA

is promoting the concept of validation, also the internal validation at universities or in companies.

ELRA has also been working on a 'universal catalogue'. ELRA's catalogue contains information about the resources provided by ELRA, whereas the universal catalogue contains information about resources identified that might be of interest to the community. The universal catalogue is at present a membership advantage.

Organizational models need to take into account that there is a cost to pay for the management of resource identification, archiving, .licensing, distribution and validation. For some resources some of these items can be free, or almost free, - e.g. the management of free resources can be dealt with in a very light way, by enabling access to free resources etc. This is one of the developments ELRA is considering.

ELRA is open to collaboration with other organizations, sharing the acquired expertise in an active partnership. E.g. a collaboration with the proposed CLARIN initiative should be explored.

5. HERA

As a very last point, we should mention the European HERA initiative (Humanities in the European Research Area). The text below is taken from the HERA project description at the EU CORDIS web site.

"During the ERA-NET Specific Support Action in 2004, The European Network of Research Councils in the Humanities (ERCH) has taken several initial steps towards large-scale cross-border coordination of research activities within the humanities. The network has now in cooperation with the European Science Foundation decided to continue the efforts under a new name: Humanities in the European Research Area (HERA). Building on the ERCH work, the HERA Coordination Action will be an extension of the network, in scope as well as depth. Firstly, the Consortium is being extended from three to fourteen members and, secondly, the range of activities is being widened to cover coordination of research activities, including the setting-up of joint research-funding initiatives. The main tasks of the CA will be:

- Consolidation of the network by establishing new network structures and integrating new members.
- Exchange of information and best practice on issues such as peer review, programme management, quality and impact assessment, and benchmarking.
- The development of research infrastructures within the humanities, which will pave the way for greater efficiency and enable new perspectives by ensuring accessibility and availability for of data and information in the widest sense.
- The ultimate objective of the CA-proposal is to coordinate research programmes in a cumulative process leading to the initiation of joint research-funding initiatives.

By applying comparative perspectives to humanities research and enabling new transnational funding schemes, it will be possible to transcend the traditional, national focus of humanities research."

It seems that it will be beneficial to explore the possibilities of cooperating with the HERA initiative, if a larger initiative covering the humanities is to be explored.

6. Acknowledgements

This paper draws upon work done in the ELRA Board, of which I am the president. In particular I want to mention contributions by Jan Odijk, Nicoletta Calzolari and Khalid Choukri.

7. References

Busa, Roberto (ed.): *Index Thomisticus,* Stuttgart, 1974, 56 volumes.

Kucera & Francis: *Computational Analysis of Present-Day American English*, Brown University Press, 1967

LDC web site, http://www.ldc.upenn.edu/

Maegaard, Bente, Khald Choukri, Nicoletta Calzolari, Jan Odijk: ELRA – European Language Resources Association – Background, Recent Developments and Future Perspectives. In: *Language Resources and Evaluation vol. 39*, Springer 2005, pp. 9-23.

McKinnon, Alastair: Computational Analysis of Kierkegaards Samlede Værker, Vol IV, Brill, Leiden, 1975