Preface

Mike Rosner
Dept. CSAI
University of Malta
MSD06
Malta
mros@cs.um.edu.mt

This workshop is something of an accident. The original intention was to have two separate workshops, one devoted to educational issues in NL technologies, and the other to infrastructures for sharing computational language resources.

Each of these proposed workshops had a history of their own.

The general problems of sharing, re-use and distribution of resources and tools has been with us for a long time. Only recently, however, have we seen the emergence of archives not only of significant size, but of a variety that reflects an amazing degree of specialisation in terms of both type of content, type of language, type of user and so on. The development of such "designer" archives is due partly to the availability of cheap computing power and mass storage (that even University laboratories can afford!), partly to a shift in perspective towards empirical methods, and partly to the ever-increasing ease with which such materials can be accessed via Internet. Good examples of existing initiatives in this area are, on the data side, ELRA/ELDA, LDC, TELRI and elsnet resource catalogues and repositories. On the tools side we have, amongst others, the ACL Natural Language Software Registry (hosted at DFKI) which was set up as a repository for tools for the distinct fields of Human Language Technology (HLT).

Mere proliferation of data is not necessarily conducive to improved practice, however, and at a workshop at ACL 2000 in Hong Kong dedicated to Infrastructures for Global Collaboration there was an agreement between the main professional organisations in NLP and Speech (ACL and ISCA), and elsnet, and the other meeting participants, that it would be useful to aim at a broadly supported, joint repository or catalogue for tools and materials for the language and speech communities. That is roughly the background to the infrastructures workshop.

The education workshop has a somewhat different history. The first point to note is that teaching in this area, whether it be from the perspective of Computational Linguistics, NLP or AI, involves a high proportion (approximately 40% according to DeSmedt's (1999) survey) of practical exercises and projects, the organisation of which depends on the availability computational infrastructure. All of those involved in teaching eventually discover the same basic problem: there are relatively few resources, both tools and data, that are *directly* useable as the basis for instruction, compared to the total volume of resources that are actually available. That is, in general, the educationally useful *yield* that results from trawling the net, is pitifully low, and may be getting lower as the supply increases. The main reason for this is that most of the material has evolved to fulfil the needs of research, not education. Consequently, the best strategy is often to just go on reinventing our own wheels for our own purposes: hardly an optimal use of existing resources.

An ELSNET-sponsored workshop was held at EACL99 in Bergen to address these issues. Some excellent individual efforts at providing tools that buck the trend were presented there (see Rosner 1999). However, the main conclusion, which, incidentally, is echoed in DeSmedt's survey, is that certain non-transient infrastructures needed to be instigated to raise the public perception of educational issues in Computational Linguistics. There was talk of founding an entity within EACL that could take on the function of coordinating training-related activities - as has already happened in ISCA. Well, this has not happened yet, but it is salutary to note that something else has: one non-transient infrastructure, founded in January 2000 under the auspices of elsnet, is the European Masters in Language and Speech, which is probably the first example of a tertiary level "qualification" that transcends national boundaries.

A second conclusion at the EACL workshop was the need for a repository of shared materials, appropriately indexed for educational usage. Some effort towards that end has been achieved through JEWELS, an EU-supported website for educational materials in Language and Speech.

It seems clear that another infrastructure that is badly needed to support such initiatives is an entity whose primary tasks would be the creation of a maintainable multilingual taxonomy of the subject matter in order to enable a consistent naming policy to be agreed upon concerning both courses, and the corresponding resources.

This workshop will build on the consensus reached at these previous workshops. As can be seen from the presentations, there are two clear foci: one upon sharable tools, the other on instruments for sharing tools and resources in general. A third theme concerns how to build upon existing initiatives as sources of data or inspiration.

The main goal of the workshop is to discuss methods for the improvement and extension of existing repositories; the educational uses of repositories; the closer interlinking between different kinds of repositories (tools and resources); global infrastructures for the achievement of joint actions. However, we expect the scope of the workshop to be much wider than that, as the issues addressed are of general interest to everybody who believes that sharing tools and resources is essential for the progress of research and education in our field.

To conclude, some words of thanks: to Thierry Declerck, Steven Krauwer and the other members of the workshop committee who all doubled as referees, for their help, patience and comments.

Our final acknowledgement goes to elsnet, but not just for sponsoring our invited speaker. If one looks carefully at many of the other initiatives that have been mentioned in this short introduction, it clear that the support shown for this workshop is no mere coincidence, but part of a dastardly long-term strategy to promote the status of our theme and also related ones within, and somewhat beyond the bounds of Computational Linguistics.

Long may this strategy continue!

References

de Smedt. K. et. al. *European Studies on Computational Linguistics*, University of Bergen 1999.

Rosner M. (ed). Proceedings of the Workshop on Computer and Internet Supported Education in Language and Speech Technology, EACL-99.

ACL NL Software Registry, http://registry.dfki.de

LDC (Linguistic Data Consortium) http://www.ldc.upenn.edu

TELRI (Trans European Language Resources Infrastructure) http://www.telri.de

ELSNET http://www.elsnet.org/resources.html

ELRA/ELDA http://www.icp.inpg.fr/ELRA

JEWELS website http://www.elsnet.org/jewels.html

.