Models of Cooperation for the Development of NLP Resources: A Comparison of Institutional Coordinated Research and Voluntary Cooperation

Oliver Streiter (1), Mathias Stuflesser (2) and Qiu Lan Weng (3)

National University of Kaohsiung (1), Department of Western Languages and Literature;
European Academy Bozen Bolzano (2,3), Institute for Specialised Communication and Multilingualism
ostreiter@nuk.edu.tw (1), mstuflesser@eurac.edu (2), steph_weng@yahoo.com (3)

Abstract

The lack of freely available resources is still the main bottleneck in the processing of the word's estimated 6000 languages. To overcome the current limitations and to achieve the impossible, one might profit from studying the models of cooperation in free software projects or Wiki-projects. In this paper, we explore such models of voluntary cooperation for the collection and elaboration of free NLP-resources. We describe the database XNLRDF which has been set up for this purpose and how data can be collected in a model of voluntary cooperation. A comparison with BLARK concludes this paper.

1. Introduction

Resources for the processing of Natural Languages are scarce. While most languages have no NLP resources at all, for those which have, the resources may be incomplete, insufficient in quality, inaccessible, very expensive or protected under copyright. The 'smaller' the language, the bigger the gaps. Proposals which intend to overcome this situation, differ in their, e.g. which languages included, their degree of concreteness, their focus, e.g. spoken vs. written language and their approach. The general failure, however, to provide a solid infrastructure for the processing of the world's languages is not as much a consequence of lacking motivation or inadequate scientific methods, as it is a consequence of inadequate models of scientific cooperation. Scientists thus should not only reflect upon how to advance science internal topics, but how to improve the architecture of science in a way that data and minds can cooperate seamlessly.

Recently, new ideas on how scientific cooperation might be organized have sprung up. These new models share the common feature of a seemingly unstructured bottom-up cooperation between knowledgeable volunteers. More and more knowledge is thus created and managed by knowledgeable volunteers in domains previously restricted to affiliated specialist. For such a cooperation, a suitable software like CVS (http://www.nongnu.org/cvs/) or Wiki is used. It provides the platform for the communication and the management of the data. Sourceforge (http://sourceforge.net) and Wikipedia (http://www.wikipedia.org) are among the flagships of this movement, and as ongoing projects show, this movement does not stop before topics related to NLP. We thus can predict that this movement will influence the field of NLP in general.

The question of quality, frequently raised to object such models, seems no longer to be a principal concern, given the acknowledged quality of free software like Firefox or the quality of the articles in Wikipedia. What's more, these models have the advantage of a potentially unlimited number of people cooperating in a flat hierarchy, creating workforces which easily superate that of the most work-intensive academic NLP projects. Overall, it seems to us that understanding the potentials of these models of cooperation will be crucial to the question whether or not we will be able to overcome the current standstill in the creation of NLP-resources.

To substantiate our claim we will first sketch the main features of traditional coordinated research on the one hand and models of voluntary cooperation on the other. Then we will present XNLRDF, a project that tries to overcome the limitations in cooperation and the standstill in the creation of NLP resources. Contrasting XNLRDF to BLARK will reveal different but complementary conceptions and priorities. The two research models thus should find a way to interact and to share their respective advantages. As volunteer communities are expanding in all directions with or without the participation of affiliated researchers, it will be up to the latter to bring the two differently powered and differently paced dynamics together.

2. Institutional Coordinated Research

Institutional coordinated research is the established model of how science is organized in universities and research centers. This model of research is stamped by (a) its funding, (b) its limitation to time slices and (c) the definition of cooperation among partners.

Coordinated research is funded by a body which, more often than not, wants its money invested in what it perceives to be relevant to the financial resources of that body. Thus, research in France, paid by French tax payers is more likely to create NLP-resources for French than for Khamtanga. This creates a distorted relation between requirements and funding. As a consequence of this ego-centrism of the funding bodies, those languages, which have the smallest gaps in their infrastructure, receive most funding.

The second feature of institutional coordinated research is it's temporal limitation. Money is granted in well defined projects with starting and end points more likely to be determined by the accounting scheme of the funding body than related to inherent features of the project. Although this temporal limitation seems to be an organizational necessity, it is also obvious that topical or relevant research is an activity that should have started before the project and should continue after the project, otherwise research islands will be financed.

Finally, the cooperation between research units (e.g. univiersities or research centers) is organized in modules where the participating research units are autonomous within their modules and interact with other modules through specific interfaces, standards or protocols. In this model of cooperation, intellectual properties can be easily assigned to the research units. In addition, the consistency and coherence of the data within one module seem to be manageable. However, this model cannot take direct advantage of closely overlapping intellectual competences where a minor gap in one unit can be easily filled by another unit.

3. Models of Voluntary Cooperation

Funding is not a condition for a project to start. Volunteers cooperate on the realization of a content, be it software, lingware, translation, images or a new texts, despite the absence of any funding organization (c.f Bey et al. 2005). The only criteria for setting the research topic is the perceived relevance by the volunteers who, although not free of a ego-centric perception, can accommodate more easily to unbiased views. Thus, while in the institutional cooperation, no language resources are created for Khamtanga, except in Ethiopia itself, researchers from France and many other countries would contribute to the development of Khamtanga NLP-resource in the model of voluntary cooperation.

The shelf-life of such projects is much longer than the time span granted to institutional project, up to a decade or more. Although volunteers may drop out of a project defined within a voluntary cooperation, such projects tend to continue as long as the development seems to be relevant to the community. Thinothings temporal continuity is one of the principal assets which put such projects and not institutional project in state to improve, for example, the infrastructure for the processing of the world's languages.

Finally, the cooperation in these projects is less likely to be modular. This feature thus allows different people to work on the same data, so that overlapping but different knowledge resources can be merged. For this reason, this cooperation requires the support of a software which tries to minimize the friction between the cooperating units. A number of complementary techniques are used at these aims, depending mainly on the degree to which data are formally structured. A common feature is the fragmentation of data (e.g. in paragraphs in Wikipedia), so that each fragment can be worked on individually. Especially powerful are relational databases which provide maximal fragmentation. The overall coherence and consistency can be controlled through uniqueness constraints, references and triggers. Backup functions and human control complete the set of control mechanisms. Linking such a database to a web-interface, allows a person connected to the Internet to cooperate almost instantly without setting up a project design or a software.

4. XNLRDF, Exploring Voluntary Cooperation

In order to bridge the gaps between a) the needs of languages users, b) affiliated research and c) the potential contributions of non-affiliated voluntary researchers, we started to create an environment for the cooperation through the Internet with the aim to collect and elaborate NLP-resources for the world's writing systems (Streiter & Stuflesser 2005). Simply the scope of the project, with currently 23.000 writing systems, 8000 languages, textual examples of 600 writing systems and 150 scripts, makes it obvious that models of traditional institutional cooperation do not provide a straight solution for the management of such an endeavor. Funding of such a project by any national body is unlikely given its multi-national/multi-regional scope. In addition, the required intellectual contribution (e.g. language expertise and access to resources) is beyond what a group of people can achieve in their life-time. Therefore, models of collaborative work are explored to achieve a Wikipedia-like cooperation of researchers.

Currently, the discussion of data structures and the collection of the first data is done by a small circle of volunteers and a few affiliated scientists. Within one year however, we succeeded in creating a basic architecture for the development of fundamental NLP-resources and to populate the database with data from the writing systems of the world. The potential of these resources starts to get visible with an automatic language recognizer and a basic spell checker working for currently about 600 languages.

The created NLP-resources are available in hourly builds under the GNU public license at http://140.127.211.214/research/nlrdf_download.html and intellectual insights related to the development of the resources are available under the Creative Commons License at http://xnlrdf.wikispaces.com. We hope that the circle of interested people might gradually enlarge, to open up finally for a free Wikipedia-like cooperation.

To prepare the project for a larger group of cooperators a number of design features have been proposed and are currently being implemented.

4.1 Relational database and XML

Not XML, but a relational database serves as backbone for data development. Modifications in the relational database affect individual data cells and thus provide a maximal fragmentation. This contrasts with XML which is normally organized in large text files.

Within the project, XML is used only for the exchange of data in RDF (cf. Manola and Miller 2004), hence the name XNLRDF. The database dump and a one-to-one representation of dump in XML can be already downloaded. An RDF will be designed which, due to the size of the download, will allow for extracts for single languages/writing systems.

The relational database has the advantage to integrate a client-server architecture for global collaborative work.

In addition to the standard clients for the management of the database and the data, there has been created a web interface that will allow volunteers to enter and check data. This interface, called the XNLRDF-browser, is accessible under http://140.127.211.214/cgi-bin/gz-cgi/browse.pl and allows its user to get an insight into the nature and wealth of the data. While currently the database still requires passwords for data modification, the system is gradually being opened up for voluntary collaboration as the appropriate checks have been tested.

Through the integration of simple tools into the XNLRDF-browser, which among others test the potential of XNLRDF, people should become motivated to enter data, e.g. to insert open licence-texts for a language to download seconds later a simple spell-checker.

4.2 Assuring quality

Quality checks are being installed at three levels: at database management level, at expert validation level and at the level of the voluntary user interface.

4.2.1 Coherence and consistency checks

Checks of coherence and consistency are being installed at the level of the database itself. These checks will thus apply to all kinds of servers that connect to the database. The checks can be defined to any level of complexity using triggers and functions.

Creating ambiguous data becomes impossible through uniqueness constraints.

References make it impossible to delete central data, e.g. a language referred to by a writing system.

The inclusion of false positives, e.g. pejorative language names, marked as deleted, makes it impossible to insert or inherit the same value again through the effect of uniqueness constraints.

4.2.2 Expert validation of data

The linguistic validity of inserted data is checked at the second level by a group of registered experts which delete unwanted data, freeze good data and wait for other data to be improved. At this level, the experts have to be informed on who created which data and how the data relate to the other data. Expert valdiation is to be done by two kinds of experts: by proficient speakers and by experts in formal or computational linguistics.

Letting linguists and native speaker experts declare an ever growing number of data in a network of data as unchangeable (freezing) will make the space for incorrect modifications smaller and smaller.

4.2.3 Guidance of voluntary users

At the level of the interface, voluntary users are guided to create as valid data as possible. On the one hand, this is done by making data fields obligatory and using picklists instead of free input. On the other hand, the database shows immediately the results of insertions, and the volunteers can see and check their input. Additionally, the setting up of the helpfile page for XNLRDF (http://xnlrdf.wikispaces.com/) has contributed to the usability of the database.

5. BLARK

BLARK, the Basic LAnguage Resource Kit tries to initiate coordinated actions to fill the gaps observed in the infrastructure for the processing of European languages. The project's aim is thus fully compatible with that of XNLRDF. The formal limitation of BLARK to European languages is not an inherent feature of the project but maybe necessary only to become 'fundable' and thus realizable in a traditional research framework. In fact, a BLARK-matrix for Arabic has been created (www.nemlar.org) beside BLARKs for European languages.

The idea of BLARK has been born in the Netherlands (Krauwer 1998), to be proposed as a project for the Fifth Framework Program of the European Commission. Unlike a project of voluntary cooperation this project idea didn't take off before any official funding was obtained. Revitalized in the Enabler Network, the concept gained the status of a reference according to which the development of NLP resources for a language can be measured and actions can be motivated. The amount of new language data however created under the direct influence of BLARK are relatively limited. Thus, although the Enabler Network shares our criticism of current funding policies and their incapacity to provide NLP-resources, the products of the network are basically theoretical in nature. Depending on national funding, BLARK thus lead mainly in the Netherlands and France to the production of new NLP resources (Mapelli & Choukri 2003). This minor impact on not more than 0.025% of the languages, falls back behind the ultimate goal of XNLRDF and BLARK which is to provide NLP resources to a great number of languages.

6. Conclusion

Thus, although BLARK and XNLRDF have similar goals and share a similar skepticism concerning the potential of current funding schemes to overcome the general misery in language resources, the contribution of BLARK is mainly in the development of concepts, schemes and metadata. Actually, nothing else could be expected, as the traditional models of scientific cooperation are not overcome but repeated within the project. XNLRDF however, without any funding and without a strong theoretical overhead, created very elementary language data (currently texts and wordlists, list of number words, function words etc) and very elementary tools (Liu et al. 2006) for about 600 languages through the participation of volunteers within one year. It is thus more than obvious, that bringing together the two research models, institutional coordinated research and voluntary cooperation, in one cooperative project design would provide the volunteers with sound conceptual features, whereas volunteers and language activists are willing and capable to provide urgently needed language data, especially for languages that start to be explored in electronic media.

To sum up, using BLARK and XNLRDF as examples, we tried to show that traditional models of scientific cooperation are unlikely to bridge the most urgent gaps in the development of NLP resources, while models of volunteer cooperation have the potential to do so. Researchers acquainted only with the first model of research should become aware of the enormous potential in voluntary cooperation and try to bring the two traditions together, profiting from the two motors the two traditions are powered by. Funding organizations should also acknowledge the potential of voluntary cooperation and support frameworks in which both traditions are alive.

7. References

Bey Y., Kageura K. & Boitet. Ch. (2005). A Framework for Data Management for the Online Volunteer

Translators' Aid System QRLex. In *Proceedings of 'PACLIC 19, the 19th Asia-Pacific Conference on Language, Information and Computation'*, 1st-3rd December 2005, Taipei, Taiwan.

Krauwer, S. (1998). ELSNET and ELRA: A common past and a common future, in *ELRA Newsletter* Vol. 3 N. 2. 1998.

Liu, D., Su, S., Lai L, Sung, E., Hsu, J. & Hsieh, S. & Streiter, O. Collaborative Development of African Language Resources. In *Proceedings of 'Networking the development of language resources for African languages'*. LREC Workshop Genoa, Italy, 22 May 2006.

Mapelli, V. & Choukri, Kh. (2003). ENABLER, European National Activities for Basic Language Resources, Deliverable D5.1, Report on a (minimal) set of LRs to be made available for as many languages as possible, and map of the actual gaps, June 2003.

Streiter, O. & Stuflesser, M. (2005). XNLRDF, the Open Source Framework for Multilingual Computing, in: I. Ties (ed.) *Lesser Used Languages & Computer Linguistics*, European Academy Bozen Bolzano, Italy, 27th-28th October 2005. http://140.127.211.214/publs/files/nlrdf_lulcl_10.pdf