Building Distributed Language Resources by Grid Computing

Fabio Tamburini

University of Bologna - Italy
f.tamburini@cilta.unibo.it

Abstract

The increasing demand for linguistic resources consisting of substantial amounts of data, such as large corpora, presents the challenge of building computational infrastructures capable of handling unprecedented amounts of information. One possible solution is the sharing of high-level, linguistically motivated and carefully balanced corpora for building one large language resource accessible worldwide. The most feasible way of integrating such widely distributed resources seems to be the construction of an infrastructure to connect various sites by interfacing local presentation formats, access methods and policies in a global network to automatically manage access procedures to widely distributed and diversified materials. Grid computing systems are designed to meet these requirements. This paper presents work in progress on an experiment for building a distributed corpus structure prototype. A small web portal was designed to perform global queries in the distributed corpus and collect the results of the same query applied to each local corpus forming part of the grid. Moreover, other computational services such as an online POS tagger and a morphological analyser/generator were inserted into the Grid to show the feasibility of such scenario.

Introduction

In recent decades the research in linguistic fields has radically changed perspective, making an increasing use of empirical evidence derived from authentic data. The demand for huge linguistic resources, such as large corpora, has increased tremendously both for language analysis and for collecting statistical data able to drive the development of powerful and affordable Natural Language Processing (NLP) applications. This increasing demand for linguistic resources consisting of substantial amounts of data presents the challenge of building computational infrastructures capable of handling unprecedented amounts of information.

Corpora size has now reached the GigaWord frontier, especially if we consider the potential for constructing corpora by pooling the different resources spread across the globe. However, the proper management of a 1GWord corpus is a demanding task, both in terms of technological infrastructure and human resources. The document retrieval and formatting, corpus management, and software design and maintenance for such large amounts of data are all extremely demanding tasks, difficult to perform in a satisfactory way in one single organisation.

One solution to the need for huge amounts of data has been proposed by the WebCorpus team (Kilgarriff, 2001) but, as some scholars have pointed out (Klein, *et al*. 2003), the use of raw documents from the Internet presents major drawbacks and is linguistically noisy, especially in terms of representativeness. Due to the extremely heterogeneous nature of online documents, the linguistic accuracy of this amount of raw data is generally poor. Moreover, the widespread presence of scripts and programs in Internet sources can corrupt the reliability of query results and affect the global statistics derived from them.

Many scholars (Biber, 1993; Váradi, 2001) have pointed out that the representativeness of the source data matters both in terms of linguistic inquiry and in terms of statistical information. The Internet as a whole does not seem to provide a good sample of actual language in use, even if the extremely high quantity of available data could, in principle, hide or reduce the bias introduced by such phenomena in language investigation.

An alternative solution is the sharing of high-level, linguistically motivated and carefully balanced corpora for building one large language resource accessible worldwide. There are a lot of widespread resources that were built after a careful linguistically-justified design phase that can provide large quantities of balanced data for the studies both linguistically oriented or NLP oriented. Such corpora are often built in conjunction with publishers or other commercial partners; the use of raw material derived from commercially distributed documents leads to the construction of resources covered by copyright, not allowing for distribution or the construction of one huge corpus resource located in a specific place under the control of just one organisation**.**

From the above considerations it seems that the barrier of 1GW corpus built in a reliable way cannot be overcome, at least in the perspective of a worldwide distribution serving different countries, organisations and scholars.

Some researchers (Hughes, Bird, 2004; Klein, *et al*. 2003) have suggested a possible solution for these problems in the perspective of global resource sharing. The most feasible way of integrating such widely distributed resources seems to be the construction of an interconnection to link all these sites with a global infrastructure to provide access to these resources in a controlled, secure and distributed way. This infrastructure has to allow interfacing with local presentation formats, access methods and policies in a global network and automatically manage procedures for accessing widely distributed and diversified materials. The infrastructure needs to exchange the required data in a coherent and transparent way, giving the user the impression of accessing a single coherent large corpus, while maintaining control of the local resources by the host institutions and enabling authorised users to access them in a secure way.

The general picture that emerges from previous considerations fits quite well into the computational schemata of Grid technology (Leinberger, Kumar, 1999; Baker, *et al.* 2000; Foster, Kesselman, 1999; Foster, *et al*. 2001; Krauter, *et al*. 2001). According to Foster and

colleagues, the idea underlying the Grid is "*coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organisations*"; data, formats, software and computational/linguistic resources are shared between all the partners taking part in the Grid. Grid computing systems are designed to meet the needs of linguist-community resource sharing. Support for Grid technology consists of a suite of tools connecting different computing infrastructures together. Access to the Grid enables the user to execute commands on remote machines, search databases and exchange large amounts of data in a controlled and secure manner.

Other disciplines such as high-energy physics (Avery, 2003), brain science (Buyya, *et al*. 2004), drug design (Buyya, *et al*. 2003) and many others have already taken advantage of such technologies, building large Grids to connect dozens of sparse research sites in one virtual organisation.

One of the leading suites currently available for building a Grid is the Globus Toolkit (Foster, Kesselman, 1997). It provides some basic services for building the fundamental structure of the Grid and a lot of additional services for managing specialised applications or setting up advanced functions**.**

The main elements composing the suite are:

- GRAM (Globus Resource Allocation Manager)
  A base service providing capabilities for remote job submission and control. The authorised user can submit a job to a remote computing resource in a standardised way without having to pay attention to specific commands/procedure to apply in the remote site.
- GSI (Grid Security Infrastructure)
  A service providing basic security controls across the Grid. Every operation both local and remote that uses the Grid infrastructure support is verified and user credentials are checked.
- GASS (Globus Access to Secondary Storage)
  With GASS services the user can move large amounts of data across the net using the secure infrastructure provided by the Grid services.
- GIS (Grid Information Service)
  Also known as Metacomputing Directory Service, it provides information services across the Grid. The user can query this service to obtain information about the available facilities as well as the computing capabilities, available linguistic resources, network bandwidth, etc.

The focus will now be on experimental work in progress at CILTA – University of Bologna – for building a distributed archive of language resources and computational procedures for performing language processing tasks using the services provided by the Grid in a small and controlled environment.

The Grid

The small-scale experiment for the construction of a computational Grid able to accommodate access to distributed language resources performed at CILTA is designed to test the feasibility of using such a technical infrastructure for the needs of natural language processing.

Three main services are included in the experimental Grid, all referred to Italian language:

- a 100Mw corpus, namely CORIS/CODIS (Rossini Favretti, *et al*. 2002), distributed across the Grid;
- a POS tagger (Tamburini, 2000);
- a morphological analyser/generator (Battista, Pirrelli, 1996a, 1996b).

The CORIS corpus was split into three slices stored on separate computing nodes on the grid to simulate three different Italian corpora joined in a unique resource. Each node has been provided with a software module enabling remote access to the local corpus data. The approximate sizes of the three slices were 63 Mw, 22 Mw and 15 Mw.

The other two computational tools, the POS tagger and the morphological analyser/generator, have been installed on separate computational nodes, both accessible remotely using appropriate software modules through the Globus Toolkit facilities.

Figure 1 illustrates the overall design of the Grid set up for this experiment.

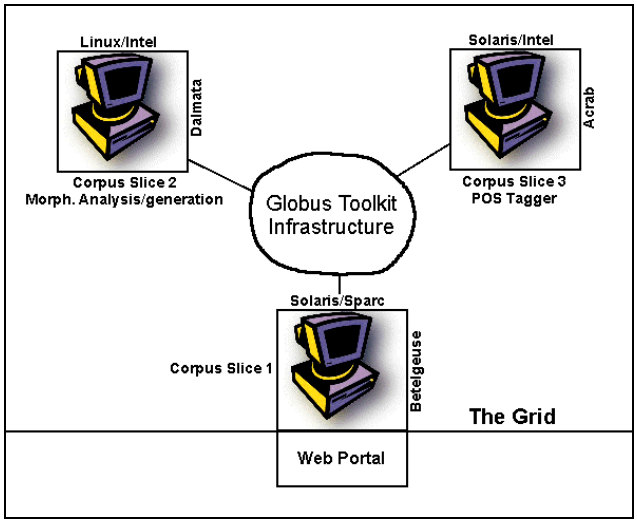

Figure 1: Overall design of the Grid set up for this experiment.

The three computational nodes composing the Grid are heterogeneous systems to test the effectiveness of the Globus support across different architectures, though all the nodes are Unix machines.

It is common practice to create a Web portal for accessing Grid services, both for simplifying the use of the different available resources and managing all the structural differences presented by the various resources in a controlled and centralised way. In this way the user does not need to know the protocol for accessing a specific resource. The portal is responsible for the proper formatting of the data (both in the input stream and in the output stream), for dispatching the computation or the information retrieval task to the proper computational node and for collecting the results and assembling them in the correct way for the user.

Following this philosophy a small web portal was designed to perform global queries to the distributed corpus and collect the results of the same query applied to each local corpus forming the grid (see figure 2).

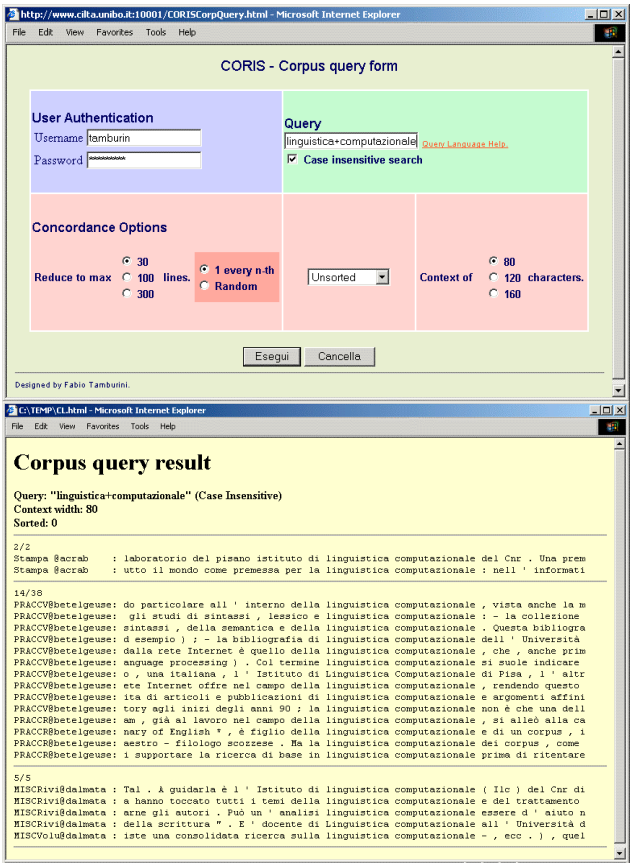

Figure 2: Distributed corpus querying. Note that the concordances belong to the three different computational nodes involved in the Grid (namely, Acrab, Betelgeuse and Dalmata).

The other services, namely POS tagging and morphological analysis/generation, are accessed in the same way by using the web interface provided by the portal. With regard to POS tagging, the user places the text to be tagged in a box and, once submitted, the system transfers the input data to the computational node hosting the tagging service through the GASS facilities provided by the Globus Toolkit. It then remotely executes the tagging process and retrieves the output data displayed on the screen (see figure 3). Considering the morphological analyser/generator service the task is easier: having inserted the word-form/lemma into the web form and selected the operation requested, the system queries the remote mophological module to obtain the proper output and display it. Figure 4 shows the generation process starting from the lemma "*bicicletta*".

All the interactions between the nodes composing the Grid are managed and integrated with the GSI module, to guarantee secure authentication and data transfer in every step of resource access. The easiest way to set up such an environment is to create a Private Certification Authority inside the Grid Virtual Organisation and manage all the credentials involved in the process in a centralised way.

As shown in Figures 1 and 2**,** the user access is now authenticated through password checking. A further improvement involves the use of certificates also on the user side and not only for authenticating the different

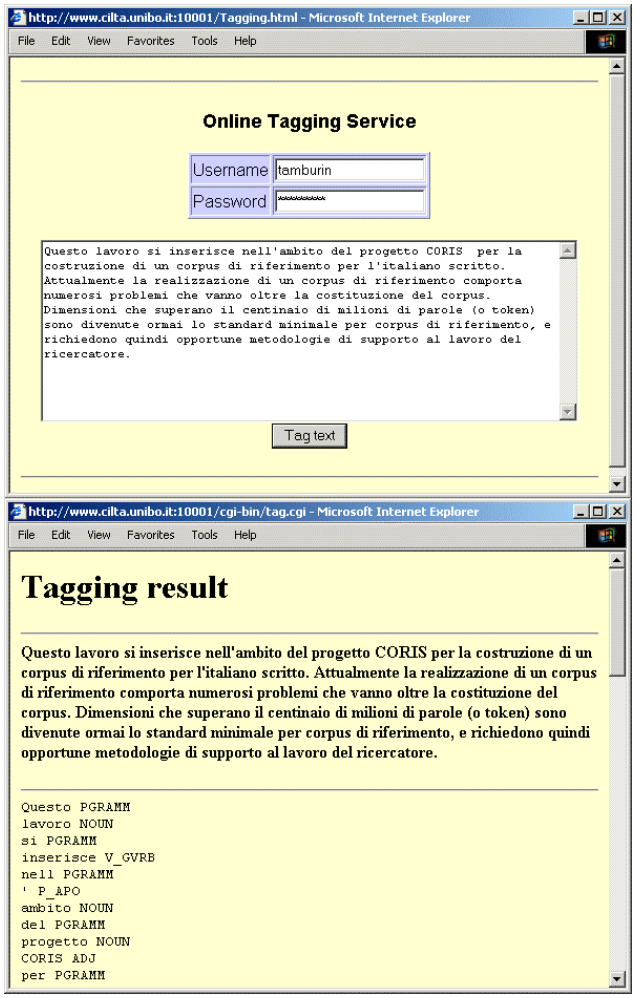

Figure 3: The tagging process.

blocks composing the Grid. This option is currently under investigation.

Conclusions

This paper presented a small-scale experiment for testing the feasibility of the construction of a Grid infrastructure for Natural Language Processing. Some data resources and computational tasks were distributed across the nodes connected to the Grid and linked using the Globus Toolkit to form a unique language resource repository. Moreover, a small Web portal was built to provide easy access to such resources. The Grid designed, though small and prototypical, shows that sparse linguistic resources can be integrated in a worldwide common repository, leaving ownership and control of the local resources with the respective owner.

It is important to note that the main contribution of this experiment is in the integration of different resources; querying corpora using the Web, as well as accessing POS tagger or morphological modules over the Internet is not new. The contribution of the present experiment is mainly concentrated on the integration of such resources, spatially distributed and potentially owned by different organisations, in a unique repository that allows the enabled user to access a potentially infinite set of language resources, both for research and teaching purposes. Although the experiment was performed in a restricted Local-Area-Network domain, the same global

infrastructure can be maintained unchanged also for Grids distributed worldwide.

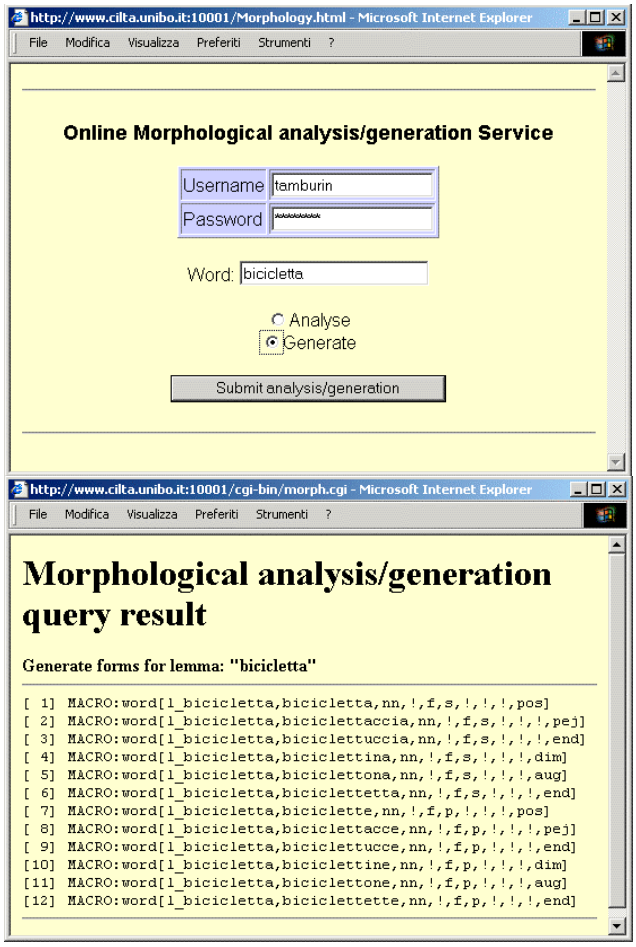

Figure 4: The word-form generation process.

A common XML-based data format for queries and results is currently at the design stage for use for each interaction between the servers connected via the Grid infrastructure. The definition of such interaction schemata will make it possible to hide completely the formatting, software behaviour and specific access procedures of the local resources**,** enabling the use of that resource simply by interfacing the local specific procedures with a small piece of code for encapsulating the results in a standardised XML data format common to all the shared resources.

References

Avery, P. (2003). Grid Computing in High-Energy Physics, In *Proc. of 9th International Conference on B-Physics at Hadron Machines – BEAUTY 2003*, Pittsburgh.

Baker, M., Buyya, R. and Laforenza, D. (2000). The Grid: International Efforts in Global Computing, *Software – Practice and Experience*, 32 (15), 1437-1466.

Battista M. and Pirrelli V. (1996a), Monotonic Paradigmatic Schemata in Italian Verb Inflexion, In *Proc. of COLING-96*, Copenhagen, 77-82.

Battista M. and Pirrelli V. (1996b), Descriptive Economy and the Morphology Lexicon, In *Proc. of EURALEX-96*, Göteborg.

Biber, D. (1993). Representativeness in corpus design. In *Literary and Linguistic Computing*, 8 (4), 243-257.

Buyya, R., Branson, K., Giddy, J. and Abramson, D. (2003). The Virtual Laboratory: a toolset to enable distributed molecular modelling for drug design on the Woorld-Wide Grid. *Journal of Concurrency and Computation: Practice and Experience*, 15, 1-25.

Buyya, R., Date, S. Mizuno-Matsumoto, Y., Venugopal, S. and Abramson, D. (2004). Neuroscience Instrumentation and Distributed Analysis of Brain Activity Data: A Case for eScience of Global Grids, *Journal of Concurrency and Computation: Practice and Experience*, to appear.

Foster, I. and Kesselman, C. (1997). "Globus: A Metacomputing Infrastructure Toolkit". *International Journal of Supercomputer Applications*, 11 (2), 115-128.

Foster, I. and Kesselman, C. (eds.) (1999). *The Grid: Blueprint for a New Computing Infrastructure*, San Mateo: Morgan Kaufmann.

Foster, I., Kesselman, C. and Tuecke, S. (2001). The Anatomy of the Grid: Enabling Scalable Virtual Organisations. *International Journal of Supercomputer Applications*, 15 (3), 200-222.

Hughes B. and Bird, S. (2004). A Grid-Based Architecture for High-Performance NLP**.** *Natural Language Engineering*, to appear in June 2004.

Kilgarriff, A. (2001). Web as corpus. In *Proc. of Corpus Linguistics 2001*, Lancaster.

Klein, E., Osborne, M. and Holt, L. (2003). GridNLP. http://www.ltg.ed.ac.uk/~ewan/WIP/

Krauter, K., Buyya, R. and Maheswaran, M. (2001). A Taxonomy and Survey of Grid Resource Managment System for Distributed Computing, *Software – Practice and Experience*, 32, 135-164.

Leinberger, W. and Kumar, V. (1999). Information Power Grid: The new frontier in parallel computing? *IEEE Concurrency*, 7 (4), 75-84.

Rossini Favretti R., Tamburini F. and De Santis C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In Wilson, A., Rayson, P. and McEnery, T. (eds.), *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, Munich: Lincom-Europa.

Tamburini F. (2000). Annotazione grammaticale e lemmatizzazione di corpora in italiano. In Rossini Favretti R. (ed.), *Linguistica e informatica: multimedialita', corpora e percorsi di apprendimento* (pp. 57-73), Roma: Bulzoni.

Váradi, T. (2001). The linguistic relevance of Corpus Linguistics. In *Proc. of the Corpus Linguistics 2001 Conference - CL2001* (pp. 587-593). Lancaster University.