

# A Non-Profit Operation Model for the Language Grid

Toru Ishida<sup>1,2</sup>, Akiyo Nadamoto<sup>2</sup>, Yohei Murakami<sup>2</sup>, Rieko Inaba<sup>2</sup>, Tomohiro Shigenobu<sup>2</sup>,  
Shigeo Matsubara<sup>1,2</sup>, Hiromitsu Hattori<sup>1</sup>, Yoko Kubota<sup>1</sup>,  
Takao Nakaguchi<sup>3</sup>, and Eri Tsunokawa<sup>3</sup>

<sup>1</sup>Department of Informatics, Kyoto University, Japan  
{ishida, matsubara, hatto, yoko}@i.kyoto-u.ac.jp

<sup>2</sup>Language Grid Project, National Institute of Information and Communications Technology, Japan  
{nadamoto, yohei, rieko.inaba, shigenobu}@nict.go.jp

<sup>3</sup>NTT Advanced Technology Corporation, Japan  
{takao.nakaguchi, eri.tsunokawa}@ntt-at.co.jp

## Abstract

The Language Grid is an initiative to build an infrastructure that allows end users to create new language services for their intercultural / multilingual activities. To this end, language resources (including data and programs) are wrapped to form Web services that users can combine easily to realize workflows that suit their own activities. There are four types of stakeholders in the Language Grid: *Language Resource Provider*, *Computation Resource Provider*, *Language Service User*, and *Language Grid Operator* (who coordinates the other stakeholders). Though there can be various operation models for the Language Grid, we propose a non-profit operation model in this paper. This model limits the usage of language services solely to non-profit operations, tries to match the incentives of stakeholders, and manages various problems associated with intellectual property rights, user privacy, and operation costs.

## 1 Introduction

In 2002, we conducted a six month experiment called the *Intercultural Collaboration Experiment (ICE2002)* to develop open source software with Chinese, Korean and Malaysian colleagues (Nomura *et al.*, 2003). We thought that machine translation would be useful in facilitating intercultural activities. We gathered machine translators to cover five languages: Chinese, Japanese, Korean, Malay and English. More than forty students and faculty members joined this experiment during

which they discovered the accessibility and usability of machine translators.

Even though there are many language resources (both data and programs) on the Internet (Choukri, 2004), many ongoing intercultural collaboration activities are still lacking multilingual support. We have been working with nonprofit organizations to investigate the requirements placed on the Language Grid. One NPO<sup>1</sup> have been creating a “universal playground” for kids around the world. To plan activities among Japan, Korea, Kenya and Austria, volunteer facilitators had to use their own dictionaries for performing translations. Another NPO<sup>2</sup> assisting foreign patients must use multilingual parallel texts for medical support. However, most end users have no way of employing the extant language resources, because of complex intellectual property rights, non-standardized technical interfaces, and so on. If technologies were available that could provide software to coordinate stakeholders, to combine language resources, and to create language services for end users, it is likely that people will start to use language resources in daily life (Ishida *et al.*, 2007).

The Language Grid project (Ishida, 2003) aims at wrapping language resources as Web services so as to make it easy to manage intellectual property. *Language Grid Users* can take three different roles: *Language Resource Provider*, *Computation Resource Provider*, and *Language Service User*. To conclude agreements between Language Grid

<sup>1</sup> NPO Pangaea. <http://www.pangaean.org/>.

<sup>2</sup> Kyoto Center for Multicultural Society  
<http://www.tabunka.jp/kyoto/>.

Operator and Language Grid User, we need to discuss operation models of the Language Grid<sup>3</sup>.

Though various operation models are possible for the Language Grid, this paper proposes a *non-profit operation model*, which limits the usage of language services to non-profit activities. The model is designed to fulfill the requirements of all stakeholders. Because the model is intended for non-profit use, universities or research institutes should be able to operate the Language Grid, and thus the cost of operation should not be too large. The rest of paper describes the service layer, stakeholders, requirements, and proposed operation model of the Language Grid.

## 2 Service Layer

### 2.1 Service Layer of the Language Grid

As shown in Figure 1, the Language Grid consists of four service layers.

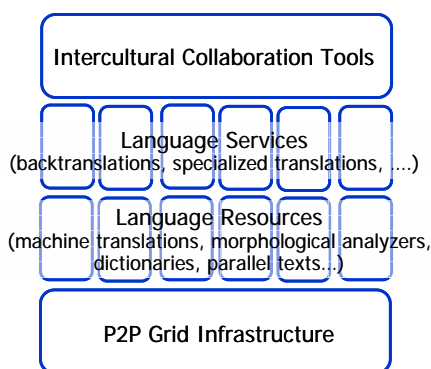


Figure 1. Service Layer

- a) Intercultural Collaboration Tools  
Collaboration tools are developed using language services explained below. Though the Language Grid provides several tools, new tools can be developed by Language Service Users, and also existing tools can be multilingualized.
- b) Language Services

<sup>3</sup> Besides language and computation resource providers, we need to consider providers of Web service wrappers, Web service workflows, and various multilingual collaboration tools that use the Language Grid. To simplify the discussions, we do not explicitly mention those stakeholders in this paper. We assume that the above software, as well as Language Grid software that run on Language Grid servers, are already licensed to the Language Grid Users.

Various language services will be available by combining existing language resources. We already implement Web service workflows including *back translations*<sup>4</sup> and *domain specific translations*. Language Service Users can easily add new language services to the Language Grid by themselves.

- c) Language Resources  
Various language resources will be provided as atomic Web services on a standardized interface. Language Resource Providers can easily add new language resources to the Language Grid, and access the usage statistics of their resources.
- d) P2P Grid Infrastructure  
This infrastructure organizes multiple servers on the Internet to fulfill end users' requests. Computation Resource Providers can add their servers to the P2P Grid, and access the usage statistics of their resources.

### 2.2 Intercultural Collaboration Tools

The Language Grid provides several multilingual communication tools that combine community dictionaries and machine translators.

- a) Langrid Input  
A multilingual input interface for existing collaboration tools. As shown in Figure 2, input texts are translated, in real time, into various languages and sent to collaboration tools. Language Service Users can multilingualize existing tools by attaching Langrid Input interfaces to them. Users can also edit dictionaries of community vocabularies<sup>5</sup>.
- b) Langrid Chat  
A chat tool with multilingual translations. Users can read and write messages in their first language. Langrid Input is used as a text input interface.
- c) Langrid Blackboard  
A tool for summarizing and sharing multilingual information. Users can create cards in their first languages and post them in a shared workspace. The texts on the cards are immedi-

<sup>4</sup> To allow users to check the accuracy of translation, collaboration tools often provide *back translation*, which translates the translated sentences into the original input language.

<sup>5</sup> Natural Disaster Youth Summit Committee has been using NGO iEARN's BBS with the Langrid Input. <http://ndys.jearn.jp/>

ately translated into different languages. This tool is useful for international meetings. Langrid Input is used as a text input interface<sup>6</sup>.

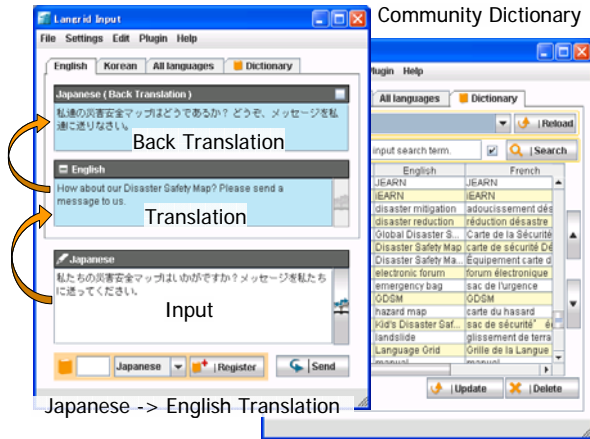


Figure 2. Langrid Input

### 2.3 Language Services

Among existing research studies, EuroWordNet (Vossen, 1998) and Global WordNet Grid (Fellbaum and Vossen, 2007) are pioneer works on connecting dictionaries in different languages based on word semantics. The Language Grid, however, is a trial to build an infrastructure that can combine language services by combining the incentives of stakeholders.

The Language Grid uses WS-BPEL to describe workflows, and uses the WS-BPEL engine to execute them; the engine sequentially invokes Web services as specified in the workflow. The following language services are available in the Language Grid:

- Atomic service: A Web service that corresponds to individual language resource. For example, bilingual dictionaries, parallel texts, morphological analyzers, and machine translators.
- Composite service: An advanced service described by a workflow that combines several atomic services. For example, multiple dictionary search, domain specific translations, and back translations.

Figure 3 shows a typical composite service: several atomic services are combined to create domain specific translation.

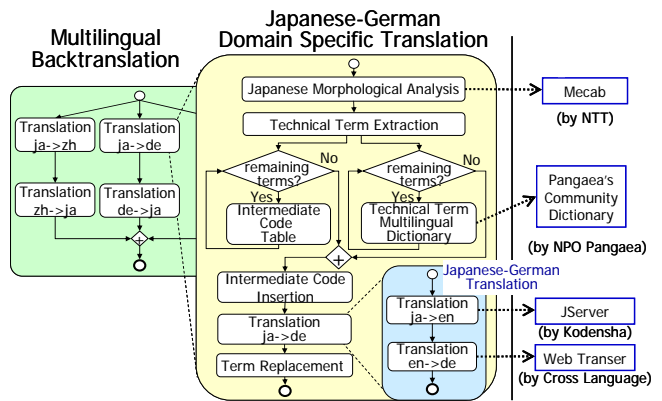


Figure 3. Example of Composite Service

### 2.4 Language Resources

The Language Grid can grow through the voluntary efforts of Language Grid Users. The more users provide resources, the more they appreciate the benefits of the resources. However, to create the initial seed, we need machine translators and morphological analyzers for both European and Asian languages.

To use language resources, we need to wrap them as Web services. Language Service Users can register Web services and share them via the Language Grid. For this purpose, standardization of access entry is quite important (Calzolari 2002). We started working on Language Service Ontology, which standardizes the interfaces for wrapping language resources (Hayashi and Ishida, 2006) (Hayashi *et al.*, 2008).

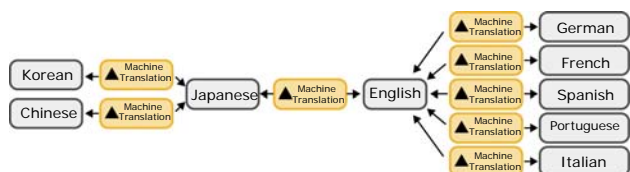


Figure 4. Available Machine Translators

To increase the flexibility of composing language services, a wide variety of language resources must be registered. However, available language resources are often limited. Figure 4 shows the machine translators available in the current Language Grid. English is often the hub of translation, and that makes back translation complex. If we use Japanese-German back translation, we have to combine four translators provided by

<sup>6</sup> Tools are available from <http://langrid.nict.go.jp/>.

different organizations<sup>7</sup>. We have observed how the quality of translation affects communication (Yamashita and Ishida, 2006). The findings have contributed to the development of new technologies to coordinate multiple machine translators.

## 2.5 P2P Grid Infrastructure

The P2P Grid Infrastructure is aimed at connecting servers around the world. As shown in Figure 5, the P2P Grid consists of two kinds of servers: *core nodes* and *service nodes*.

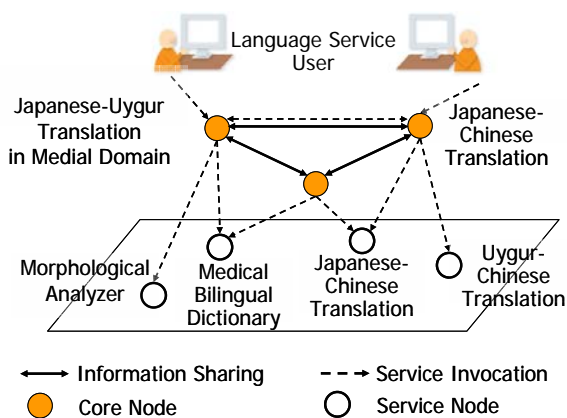


Figure 5. P2P Grid Infrastructure

Core nodes manage registered language services. They provide the search functions used by language services. Based on users' requests, core nodes invoke composite Web services by executing workflows. Information of language resources is shared among all core nodes. As a result, same services are provided, regardless of which core node is accessed by the user. The core nodes also control access to service nodes.

The service nodes deploy various language resources as Web services. The Language Grid sets basic authentication functions on the service nodes. Therefore, access from nodes other than core nodes is not accepted by service nodes.

There already exist several efforts to combine language processing programs: Heart of Gold (Callmeier et al., 2004), and UIMA (Ferrucci and Lally, 2004). Though there are similarities between Heart of Gold, UIMA and the Language Grid, their focuses are orthogonal. Heart of Gold and UIMA

<sup>7</sup> NPO Pangaea has been used Japanese-Korean and English-German translations with their own dictionaries, but does not use Japanese-German translation because of quality of back translation.

aim at allowing language processing programs with variable interfaces to share data, while the Language Grid focuses on managing the intellectual property associated with language resources (both data and programs) via the Web service architecture. We started bridging Heart of Gold and the Language Grid and will apply the results to combine UIMA and the Language Grid.

## 3 Stakeholders and Requirements

### 3.1 Stakeholders

As shown in Figure 6, the term "Language Grid User" means the three types of stakeholders.

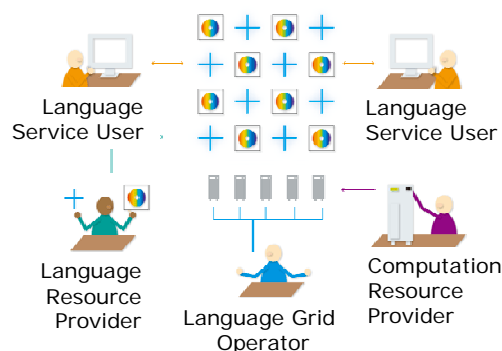


Figure 6. Stakeholders

Language Resource Providers register language resources, Computation Resource Providers register servers with the Language Grid, and Language Service Users use registered language and computation resources. Language Grid Operator concludes agreements with the Language Grid Users and manages language and computing resources.

### 3.2 Requirements

To design a non-profit operation model, we should first collect the requirements of the stakeholders. University laboratories and research institutes are expected to join the Language Grid as Language and Computation Resource Providers, and various NPOs, NGOs and public sectors are expected to join as Language Service Users. To create a non-profit operation model, we need to understand the following situations and requirements from stakeholders.

- Machine translators are often developed and operated by for-profit companies, and are provided for profit. However, if the application area of the Language Grid does not conflict with an already

existing business market, we can collaborate with those companies and receive a substantial discount on prices. One solution is that universities, research institutes or large NGOs voluntarily buy translation services and provide them to the Language Grid without any charge<sup>8</sup>.

- Morphological analyzers and other language processing programs are often developed by research institutes or universities. In many cases, researchers can provide their resources without any charge for research purposes. Even if the goal is not for research, if their use can be restricted to non-profit, researchers often agree to provide their resources. For profit tools, however, are seldom free and contracts cannot be concluded uniformly.
- Multilingual dictionaries and multilingual parallel texts may or may not be free. Even for non-profit use, if the resources are already being sold, difficult problems exist with regard to the distribution of those resources without charge. Since the Language Grid is based on Web services, however, there is a chance of making those resources freely available by setting an upper limit to daily access number.
- Most language resource providers are willing to disseminate the fruit of their research and development, expecting their resources to be widely used and to contribute to intercultural activities. However, they require that the provided resources are used properly following the agreement they signed.

The Language Grid is based on Web service technologies to combine software on the Internet. What Language Grid offers is, however, not only composite Web services but also an infrastructure wherein stakeholders can provide and/or use language resources by mutual consent, understanding and solving the intellectual property issue in each case. To meet this goal, as described in detail in Section 4, we need a *Language Grid Service Manager* that allows stakeholders to monitor the grid and to set and confirm information necessary for their participation.

We should also consider that Language Grid Operators of non-profit models are often laboratories of university or research institutes. We should

---

<sup>8</sup> At the starting point of Language Grid operation, we negotiated with commercial companies to provide the machine translators illustrated in Figure 4.

make the costs of operation as small as possible. Those Operators may not be able to handle personal information, nor invoices or transactions for the fee-based usage of language resources.

## 4 Non-Profit Operation Model

### 4.1 Language Grid Users

A Language Grid User could be more than one type of stakeholder at the same time. For example, a university that is providing a morphological analyzer might use a machine translation engine provided by another stakeholder. In this case, the university is a Language Service User as well as a Language Resource Provider. Therefore, we call all three types of stakeholders Language Grid Users. Language Grid Users are required to conclude an agreement with the Language Grid Operator.

The term “non-profit use” in this paper means the use by individuals, the use by public agencies or nonprofit organizations for their main businesses or for research, and the use by for-profit organizations for social contribution. Note that we do not exclude for-profit contracts concluded outside of the Language Grid to provide language resources with charge. What we strictly exclude is to use language resources provided via the Language Grid for profit purposes.

Language Grid Users are, in most cases, expected to be public agencies or non-profit organizations. However, private enterprises are eligible to be Language Grid Users if they use the Language Grid for voluntary social action programs, or they behave just as Language Resource Providers<sup>9</sup>. Individuals can also be Language Grid Users. For example, when a researcher working for a private enterprise provides language resources for which he/she holds the copyright, the person is eligible to become a Language Grid User.

### 4.2 Language Resource Providers

The Language Grid Operator makes language resources accessible via the Language Grid for non-profit use. The Operator can, with Providers’ consent, select or change the computation resources in which Providers’ language resources are deployed.

Language Resource Providers can set a copyright notice and/or any licensing policy informa-

---

<sup>9</sup> NTT Communication Science Laboratory provides Mecab, a Japanese morphological analyzer.

tion in the *Profile* of their language resources using the Language Grid Service Manager. Providers can also enter a URL to a Web site that describes the license.

The Language Service Users can view the profile. Furthermore, when their language service is used, the Language Grid sends copyright and license information to the Users, so that they can display the copyright and license information on their collaboration tools.

Language Resource Providers can monitor statistics on how their resources are being used. Providers can monitor the following information.

- Total number of accesses of each user to provided language services per year, month, and day.
- The registration information (name of organization, responsible person, email address, URL, etc.) of users of provided language services

Language Resource Providers can control access to the language resources as follows.

- Language Resource Providers can permit or deny access by certain users by setting access rights. When a Provider registers a language resource, its default setting is that no Language Service User is allowed to access it. The Provider then configures the settings to give permission to users. Language Resource Providers can also set the effective period of the access right.
- Language Resource Providers can restrict the total number of accesses per year, month, and day to their resource to prevent inappropriate usage. They can also set the effective period for each access frequency restriction.
- Language Resource Providers can restrict the amount of data transferred per year, month, and day to prevent the download of all data in their resources. Furthermore, they can set the effective period for data transfer. The maximum volume of data transfer per access can be set as well.

If, for some reason, Language Resource Providers want to provide their resources as paid services, they should negotiate directly with Language Service Users<sup>10</sup>. The Operator will not be involved in any contracts wherein language resources are provided with charge.

If Language Resource Providers need any detailed log information, they should negotiate di-

---

<sup>10</sup> It is prohibited to provide any language resource obtained from the Language Grid as paid services.

rectly with Language Service Users for obtaining such information. To protect the privacy of Language Service Users, Language Grid Operator will not store or collect any log information other than statistics.

If necessary, Language Resource Providers can suspend the use of their resources. Providers can stop provision of the resources by notifying the Language Grid Operator. Language Resource Providers have no responsibility for any direct or indirect damage caused by use of Providers' resources.

### 4.3 Computation Resource Providers

The Language Grid Operator makes computation resources available for non-profit purposes. The Operator can, with Providers' consent, select or change the language resources that can be deployed on Providers' computation resources.

Computation Resource Providers have rights similar to Language Resource Providers. Computation Resource Providers can obtain the registered information of Language Service Users who access their computation resources. Language Grid Service Manager allows them to monitor statistics on how their computation resources are being used. Computation Resource Providers, however, should not obtain log data other than statistical information.

If necessary, Computation Resource Providers can suspend the use of their resources. Computation Resource Providers can stop their servers without reference to the Language Grid Operator. If some server fails or is halted, the Language Grid Operator should reconfigure the computation resources dynamically. Computation Resource Providers have no responsibility for any direct or indirect damage caused by use of Providers' resources.

### 4.4 Language Service Users

Language Service Users can use language resources and computation resources, but only for non-profit purposes. A Language Service User must obtain a valid *User ID* and *Password* to use the Language Grid. Language Service Users can allow participants in events or activities organized by Language Service Users to use the Language Grid. To avoid the fraudulent usage of language resources, however, Language Service Users should not allow the participants to discover the User ID and Password. For example, in the case of an NPO offering medical-interpreter services to

foreign patients to talk to hospital doctors, the NPO should not enter their User ID and Password in front of the patients. It is also strongly recommended that Language Service Users should change Passwords periodically, and whenever they use the Language Grid in public. Furthermore, in accessing language resources, Language Service Users must conform to the terms of use expressed in each Language Resource Provider's copyright notice and/or licensing policy.

Language Service Users cannot be anonymous so as to satisfy the requirements of the Language Resource Providers. Language Service Users must agree that their statistical information will be collected and offered to the providers of the language and computation resources.

#### **4.5 Language Grid Operator**

The role of the Language Grid Operator is to conclude agreements with Language Grid Users, and to operate language and computation resources in the manner described in the agreements. Since the agreements are solely for non-profit use, the Operator is not to be involved in any for-profit contracts.

The Operator will not store or obtain any unpublished personal information. Language Grid Users are asked to input personal information when registering with the Language Grid. This information will be published on the Language Grid Operator's Web site. Also, the Operator should not obtain any usage log data other than statistics on the use of resources. All other information obtained is made public on the Operator's Web site.

Finally, the Language Grid Operator has no responsibility for any direct or indirect damage caused by use of the language and/or computation resources.

### **5 Discussion**

Below we address the questions expected to be raised by the proposed operation model.

The basic question is how the Operator can guarantee that the language resources will not be used for profit. The Operator and Users should conclude agreements to use language resources solely for non-profit, but this cannot guarantee the non-profit use of resources. In response, the Operator will conclude agreements only with trustworthy organizations and individuals. Risk of leaking the

ID and password cannot also be ignored. To prevent this, the Operator should always monitor the Language Grid. If resource abuse is detected, the Operator halts access to the language resources by the offending party.

Another question is the possibility of a large number of individuals using the Language Grid as a free resource. According to the model, non-profit use of the Language Grid includes personal use. This is because individuals may provide language resources. Therefore, Language Grid Operator will conclude agreements only with trustworthy individuals.

Can the Language Grid be used in the research labs of companies? According to the model, profit organizations can use the Language Grid only for social contribution activities, and cannot use it in their research labs. This is because language resource providers sometimes charge the research labs of companies for their resources. We need to avoid conflicts between existing businesses and the non-profit operation model of the Language Grid, which is to explore new application areas of language resources.

The non-profit operation model might be too strict to encourage the propagation of the Language Grid: Language Service Users cannot be anonymous, companies cannot use the Language Grid, most individuals are excluded in practice, only trustworthy organizations can join, and so on. Indeed, the model cuts many potential applications of the Language Grid. The non-profit operation model is mainly intended to promote the research and development of technologies with field workers. In the future, we need a sustainable model for operation, most likely some combination of for-profit and non-profit models, so that companies are encouraged to join the Language Grid, and provide high quality services to anyone in the world, just as Internet providers do at present.

### **6 Conclusion**

This paper proposes a non-profit operation model to coordinate four types of stakeholders: Language Resource Provider, Computation Resource Provider, Language Service User, and Language Grid Operator. The essence of the model is as follows.

- Language Grid Operator concludes agreements among stakeholders and monitors the Language Grid so as to maintain the agreements. The Op-

erator holds no private personal information and will not be involved in for-profit activities.

- Language Resource Providers and Computation Resource Providers can monitor the usage of their resources and control access to their resources.
- Language Service Users, which are often organizations like NPOs and NGOs, can allow participants of their events and activities to access the Language Grid.

The operation model is designed to coordinate stakeholders and to match their incentives. Trial operation of the proposed model started in December 2007 with around thirty organizations<sup>11</sup>. The Language Grid software has been developed by NICT, while its operation is done by Department of Social Informatics, Kyoto University. A user group called Language Grid Association<sup>12</sup> including NPOs, NGOs, universities, and research institutes has been formed to develop and share experiences of using the Language Grid in their activities. An analysis of the results will contribute to the increased accessibility and usability of language resources and so overcome language barriers worldwide.

## References

We give thanks to the member of Language Grid Association for their continuous collaboration. This research is partially supported by Kyoto University Global COE Program: Informatics Education and Research Center for Knowledge-Circulating Society.

## References

- U Callmeier, A Eisele, U Schafer and M Siegel. 2004. The Deep Thought Core Architecture Framework. *Proceedings of LREC*, pp.1205-1208.
- N Calzolari, A Zampolli, and A Lenci. 2002. Towards a Standard for a Multilingual Lexical Entry: The EAGLES/ISLE Initiative. *CICLing*, pp. 264-279.
- K Choukri. 2004. European Language Resources Association History and Recent Developments. *SCALLA Working Conference KC 14/20*.

C Fellbaum and P Vossen. 2007. Connecting the Universal to the Specific: Towards the Global Grid. *Intercultural Collaboration*, Lecture Notes in Computer Science 4568, pp. 1-16.

D Ferrucci and A Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, Vol. 10, pp. 327-348 Cambridge University Press.

Y Hayashi and T Ishida. 2006. A Dictionary Model for Unifying Machine Readable Dictionaries and Computational Concept Lexicons. *International Conference on Language Resources and Evaluation*, pp.1-6.

Y Hayashi, T Declerck, P Buitelaar, and M Monachini. 2008. Ontologies for a Global Language Infrastructure. The First International Conference on Global Interoperability for Language Resources.

T Ishida. 2006. Language Grid: An Infrastructure for Intercultural Collaboration. *IEEE/IPSJ Symposium on Applications and the Internet*, pp. 96-100.

T Ishida, S R Fussell and P Vossen Eds. 2007. *Intercultural Collaboration*. Lecture Notes in Computer Science, 4568, Springer-Verlag.

R Khalaf, N Mukhi, S Weerawarana. 2003. Service-Oriented Composition in BPEL4WS. *Proceedings of the World Wide Web Conference*.

S Nomura, T Ishida, N Yamashita, M Yasuoka and K Funakoshi. 2003. Open Source Software Development with Your Mother Language: Intercultural Collaboration Experiment 2002. *International Conference on Human-Computer Interaction*, Vol. 4, pp. 1163-1167, 2003.

P Vossen. 1998. Introduction to EuroWordNet. *Computers and the Humanities*, Vol. 32, No. 2-3, pp. 73-89.

N Yamashita and T Ishida. 2006. Effects of Machine Translation on Collaborative Work. *International Conference on Computer Supported Cooperative Work*, pp. 515-523.

<sup>11</sup> <http://www.langrid.org/operation/>.

<sup>12</sup> <http://www.langrid.org/association/>.