Construction of a Metadata Database for Efficient Development and Use of Language Resources

Hitomi Tohyama[†], Shunsuke Kozawa[†], Kiyotaka Uchimoto[††], Shigeki Matsubara[†] and Hitoshi Isahara[††]

[†]Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
[††]National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{hitomi, kozawa,matubara}@el.itc.nagoya-u.ac.jp, {uchimoto, isahara}@nict.go.jp

Abstract

The National Institute of Information and Communications Technology (NICT) and Nagoya University have been jointly constructing a large scale database named SHACHI by collecting detailed meta information on language resources (LRs) in Asia and Western countries, for the purpose of effectively combining LRs. The purpose of this project is to investigate languages, tag sets, and formats compiled in LRs throughout the world, to systematically store LR metadata, to create a search function for this information, and to ultimately utilize all this for a more efficient development of LRs. This metadata database contains more than 2,000 compiled LRs such as corpora, dictionaries, thesauruses and lexicons, forming a large scale metadata of LRs archive. Its metadata, an extended version of OLAC metadata set conforming to Dublin Core, which contain detailed meta information, have been collected semi-automatically. This paper explains the design and the structure of the metadata database, as well as the realization of the catalogue search tool. Additionally, the website of this database is now open to the public and accessible to all Internet users.

1. Introduction

The construction of LRs such as corpora, dictionaries, thesauruses, etc., has boomed for years throughout the world in its aim of encouraging research and development in the main media of spoken and written languages, and its importance has also been widely recognized. Of the organizations willing to store and distribute LRs, there exist some consortia fulfilling their function such as LDC[1], ELRA[2], CLARIN,[3] and OLAC[4] (Hughes et al, 2005), in Western countries, and GSK[5] which does so mainly in Japan. However, those released LRs are scarcely connected with each other because of the difference between written and spoken language as well as the difference between languages such as Japanese, English, and Chinese. This situation makes it difficult for researchers and users to find LRs which are useful for their researches. In the meantime, by connecting systematically existing various LRs with Wrapper Program, the attempt to realize multilingual translation services has already begun (Ishida et al, 2008, Hayashi et al, 2008). Moreover, since language information tags given to those LRs and their data formats are multifarious, each LR is operated individually. As LR development generally entails enormous cost, it is highly desirable that the research efficiency be enhanced by systematically combining those existing LRs altogether and extending them, which will encourage an efficient development of unprecedented LRs.

The National Institute of Information and Communications Technology (NICT) and Nagoya University, for the purpose of developing LRs efficiently, have been constructing a large scale metadata database named SHACHI[6] as their joint project by collecting detailed meta information on LRs in Western and Asian countries (Tohyama et al, 2008). This research project aims to extensively collect metadata such as tag sets, formats, and recorded contents of LRs existing at home and abroad and store them systematically. Meanwhile, we have already developed a search system of LRs by the use of meta information and are attempting the experiment of widely providing meta information on our stored LRs to those from researchers to common users. This metadata database has been now open to the public in the Web and allows every Internet user to access it for the search and read information of LRs at will. Moreover, by exchanging opinions on them with its users, efforts go into registering new LR information and appending search functions to it as required.

In this paper, Section 2 describes the purpose of the metadata database. Then, Section 3 explains the collection of metadata and the extended meta elements which characterize this database as one of its distinctive features. Section 4 discusses three search functions of SHACHI and their characteristics. Lastly, Section 5 gives an outline of statistic information about LRs registered in SHACHI.

[1]LDC :Linguistic Data Consortium, http://www.ldc.upenn.edu/.
[2] ELRA: European LRs Association, http://www.elra.info/.
[3]CLARIN: Common Language Resources and Technologies Infrastructure, http://www.ilsp.gr/clarin_eng.html.
[4]OLAC: Open Language Archives Community, http://www.language-archives.org/.
[5] GSK: *Gengo Shigen Kyokai; Language Resource Association*, http://www.gsk.or.jp/ . (in Japanese)

[6] SHACHI: Metadata Database of Language Resources SHACHI, http://shachi.org/shachi/en/index.php, (*Shachi* means “orca” in English.

2. Purpose of the construction of the metadata database

The purpose of the construction of the database is fourfold.

(1) To store language resource metadata:

It is convenient that all information about language resources in the world be put together in one place to be combined with each other by making an organic relation and to be strategically developed. Therefore, SHACHI semi-manually collects detailed metadata of language resources and constructs their detailed catalogues. Figure 1 shows a sample page of a LR catalogue stored in SHACHI (ex. Euro WordNet). This catalogue provides more detailed meta information than other LR consortia do.

(2) To systematize language resource metadata:

According to detailed metadata obtained by step (1), language resource ontology is tentatively constructed by classifying types of language resources (in this paper, it is called "ontology"). At the moment, it is under investigation what is the most useful and functional ontology for users by developing some ontologies such as human-made ontology, semi-automatically produced ontology, and automatically produced ontology. It is considered that storing systematized LRs in the world serves for the strategic development and distribution of LRs as well as the expansion of language processing technologies. Moreover, this ontology is experimentally applied to one of SHACHI search functions. (See Sec.3.3 & Sec.4)

(3) To make each language resource related to each other:

Catalogues contained in SHACHI are more characteristic than those of other consortia in terms of its collection of detailed metadata. Those detailed metadata enabled us to describe characteristics of each language resource and to expectably specify relationships among language resources. Figure 3 shows a part of the SHACHI search screen. It shows language resources found as a search result, the references to which these language resources conform as well as other language resources whose formats are common to theirs. By statistically investigating the relationships of these language resources, it is possible to design tagsets and data formats according to world level standards as well as types of language resources in demand.

(4) To statistically investigate language resources:

On the SHACHI site, users are able to peruse statistical information about language resource metadata which are contained in SHACHI. By statistically analyzing those metadata, users are able to grasp what kinds of language resources exist in different part of the world and to understand current tendencies of language resources which have been available to the public. (See Sec.5)

(5) To promote the distribution of language resources:

Since this metadata database enables users to easily gain access to language resources in accordance with their needs, owing to fully equipped search functions, SHACHI will be able to support an effective use and an efficient development of language resources. (See Sec.4)

Some 2,000 resources of metadata have already been collected in the database so far and they will be enlarged by a further 3,000 by December 2008. To that end, it is indispensable for us to work in cooperation with language resource consortia at home and abroad and to take the initiative in contributing to Asian language resources.

Consortium	ULR
Asian Language Resource Catalogue	http://nlp.kuee.kyoto-u.ac.jp/
ChineseLinguistic Data Consortium	http://www.chineseldc.org/
Corpus-linguistics	http://www.corpus-linguistics.de/
European Language Resources Association	http://www.elra.info/fr/
Global WordNet Association	http://www.globalwordnet.org/
GSK (*Gengo Shigen Kyokai*)	http://www.gsk.or.jp/
ICAME Corpus Collection	http://icame.uib.no/newcd.htm
Linguistic Data Consortium	http://www.ldc.upenn.edu/Catalog/
Natural Language Processing Portal Site	http://nlp.kuee.kyoto-u.ac.jp/NLP_Portal/lr-cat-j.html
Speech Information Technology & Industry Promotion Center	http://www.sitec.or.kr/English/
Speech Resources Consortium	http://research.nii.ac.jp/src/links/

Table 1. List of major language resource consortia which SHACHI covers

3. SHACHI metadata

3.1 Policy for collecting metadata

The LRs which our metadata database stores should satisfy the following conditions:

1. Those resources should be stored in a digital format.
2. Those resources should be one of the following: corpus, dictionary, thesaurus, or lexicon. (Numeric data are not considered to be the subject of collection for SHACHI.)
3. Those resources should be collected from English websites and its data must be open to the public.
4. Those resources should be created by research institutions, researchers, or business organizations.

It is essential to obtain and store information on highly recognized and frequently used language resources through collecting language resource metadata. Therefore, during the construction of SHACHI, a web search through Google was first conducted to investigate the actual state of language resources that meet the above mentioned conditions. LRs metadata database SHACHI contains extensive data obtained from highly recognized LRs that were searched for on the Internet, and also covers meta information provided by LR consortia such as ELRA, LDC, and OLAC whose more detailed metadata are fed into the database by semi-automatic means of importing. Table 1 shows a list of major LR consortia whose detailed LR information has already been registered in SHACHI.

		Qualifier		
		DC Qualifiers	Qualifiers used for more precise description of the resources	
	LEVEL 1		LEVEL 2	
	DCMES Eleme	DC Element Refinements	OLAC Extensions	SHACHI Extensions
1	Title	Alternative		
2	Creator	-		
3	Subject	-	Linguistic Subject (29) [olac:linguistic-field] anthropological_linguistics applied_linguistics cognitive_science computational_linguistics discourse_analysis forensic_linguistics general_linguistics historical_linguistics history_of_linguistics language_acquisition language_documentation lexicography linguistics_and_literature linguistic_theories mathematical_linguistics morphology neurolinguistics philosophy_of_language phonetics phonology pragmatics psycholinguistics semantics sociolinguistics syntax text_and_corpus_linguistics translating_and_interpreting typology writing_systems OLAC-Language extension [olac:language]	mono_multi_lingual (2) monolingual multilingual ResourceSubject (4) corpus dictionary thesaurus glossary
4	Description	Table Of Contents Abstract		Language (of description) Price
5	Publisher	-		
6	Contributor	-	Role [olac:role] (24) annotator *author compiler consultant data_inputter depositor developer editor illustrator interpreter interviewer participant performer photographer recorder researcher research_participant responder signer singer *speaker sponsor transcriber translator	Attribute of *Speaker/Author(mother-tongue intonation level age gender
7	Date	Created Valid Available Issued Modified Date Accepted Date Copyrighted Date Submitted		
8	Type	(DC Type Vocabruary)	Discourse Type (10) [olac:discourse-type] drama formulaic_discourse interactive_discourse language_play oratory narrative procedural_discourse report singing unintelligible_speech Linguistic Data Type (3) [olac:linguistic-type] lexicon primary_text language_description	Purpose(4) lexicography analysis developing_technologies education Style (2) speech written Form (2) fixed unfixed Sentence(3) short long mixed Annotation (3) annotated plain Annotaion_sample Sample
9	Format	Extent Medium		Encoding Markup Functionality
10	Identifier	- Bibliographic Citation		
11	Source	-		
12	Language	-	OLAC-Language extension [olac:language]	
13	Relation	Is Version Of Has Version Is Replaced By Replaces Is Required By Requires Is Part Of Has Part Is Referenced By References Is Format Of Has Format Conforms To		Utilization
14	Coverage	Spatial Temporal		
15	Rights	Access Rights License		

Table 2. SHACHI metadata set

SHACHI Extensions	Description
Mono_Multi_lingual (2)	Applies to: subject
monolingual	A resource using only one language: The same language is use for the subject language and to describe the subject language
multilingual	A resource in several different languages. Different language(s) are used for more than one subject languages and to describe the subject language(s).
ResourceSubject (3)	Applies to: subject
dictionary	A list of the words of a language in which the definitions or meanings of the words are explained either in the same language or in a different language.
thesaurus	A list of the words of a language in which the words are arranged in groups that have similar meanings.
lexicon	A list of words on a particular subject.
Attribute (5)	Attributed of a contributor
mother-tongue	The performer is whether a native or non-native speaker of the language.
intonation	Dialectal status whether the performer uses a standard language or a dialect.
level	Whether the performer has received a professional level of linguistic(speaking or writing) training or no such training.
age	An age group the performer is in. When there are many performers in a resource, the ratio between all the age groups to which the performers belong.
gender	Sex of the performer. When there are many performers in the resource, the ratio of males and females.
Purpose(4)	Applies to: type
lexicography	The creation of the resource is intended for lexicography.
analysis	Thecreation of the resource is intended for analysis.
developing_technolog	Thecreation of the resource is intended for developing technologies.
education	The creation of the resource is intended for the use in education.
Style (2)	
speech	The resource is of the spoken language.
written	The resource is of the written language.
Form (2)	
fixed	A collection of fixed forms of expressions.
unfixed	The resource collects various forms of expressions.
Sentence(3)	
short	A collection of short sentenses.
long	A collection of long sentenses.
mixed	A collection of varying length of sentences.
Annotation (3)	
annotated	Tagged corpus.
plain	A corpus without annotations.
annotaion_sample	A sample of tagged data
Sample	A sample of the language resource
Format (2)	
encoding	An encoded character set used by a digital resource.
markup	A markup scheme used by a digital resource.
Functionality	Software Functionality
Utilization	Applicability of the resource. The described resource is utilized for the referenced technology, education, research or a product.

Table 3. Description of SHACHI extensions

Requests for the participation in our project to research laboratories and institutes developing languages resources by registering themselves in our database are scheduled to be made, as soon as the preparations for publishing RSS feeds of our data and our input format of metadata standard are completed. At the moment our web page for registration is not open to the public.

3.2 Extensions of metadata element

Since users sometimes search for LRs without a clear objective, it is necessary for language resource providers to construct language resource ontology which enables users to find what they want by indicating the affinity of language resource attributes as well as their relationship between their classified positions in order to provide them with sufficient information. This database conforms to the OLAC metadata set which is based on 15 kinds of fundamental elements of Dublin Core and constitutes an extended vision of OLAC with 19 newly added metadata elements which were judged to be indispensable for describing characteristics of LRs. Table 2 show the SHACHI metadata set. “SHACHI Extension” in the third row is meta elements which SHACHI recently adopted and SHACHI’s originally adopted extended elements (See Table 3) are shown in the right edge of the table 2. SHACHI provides usage information about how and in which situation language resource researchers utilized each language resource, which is also important for users. In this SHACHI project, by using parsing techniques the method, through which usage information about LRs is automatically retrieved from academic article databases, is conceived and presented (Kozawa et al, 2008). (See “Utilization” in Figure1).

3.3 Systematization of LRs

The above-mentioned database aims not only to collect language resource metadata but also to store them systematically. Clear description of the relations among LRs can be applied to the efficient development of LRs and search tools for common users of database. We first surveyed the frequency of possible values of metadata element choices, obtained the standard deviation of its frequency and generated the ontology by hierarchicalizing meta elements of our meta categories in putting the meta element whose standard deviation was the smallest on the highest class. While ontology can be constructed in various ways from different standpoints, our ontology is particularly designed for users who want to search for a language resource for a certain purpose to enable to find them efficiently by following the hierarchical classes of our ontology. This ontology is applied during a trial in one of three search functions of SHACHI. Those details are described next in Section 4-(C).

4. Catalogue search tools

The metadata database is composed of language resource catalogs, a list of all language resource catalogs, a catalogue search tool by which users can retrieve the information from all LRs stored in the database from all angles, and the statistic information of the metadata of its LRs. Figure 2 shows a screen image of a search result through SHACHI.

For the purpose of facilitating users of this metadata database to find their intended language resource catalogues, SHACHI provides three search functions:

(a) Keyword search function

This tool is suitable for users who have clear images to search for specified LRs and a technical knowledge of language processing. It allows them to input keywords as they want and to search all words stored in SHACHI metadata archive. Boolean Searching is adopted by this function.

(b) Facet search function

This tool is suitable for users who have a vague idea of what kind of LR they want. It is equipped with a choice of 15 kinds of metadata elements selected from the SHACHI metadata set, from which users choose an element which seems to be the most suitable language resource for their objective. Then, they narrow down the target LRs one by one in order to find the intended one. For example, with one click on “age”, three choices such as “Children’s utterance?”, “Adults’ utterance?” and “Both are OK?” will be shown. Then, users can narrow the choice of LRs down according to their research purposes. This function can be also used together with keyword search function as the above-mentioned in (a).

(c) Ontology search function

This tool was developed by adopting the idea acquired by systematizing LRs registered in SHACHI which was described in Section 3. When using the ontology search function, users find the intended LRs by following the vertical relationship of the ontology. It was ascertained that ontology search function tool had the merit of enabling users to discover LRs that have not been ever found by keyword search and facet search functions. This fact became clear according to the experiment conducted by using 10 subjects who were all research workers in information processing or language processing fields. It was also revealed that this tool could indicate appropriate LRs more effectively than other two functions could. This function can be also used together with keyword search function as the above-mentioned in (a).

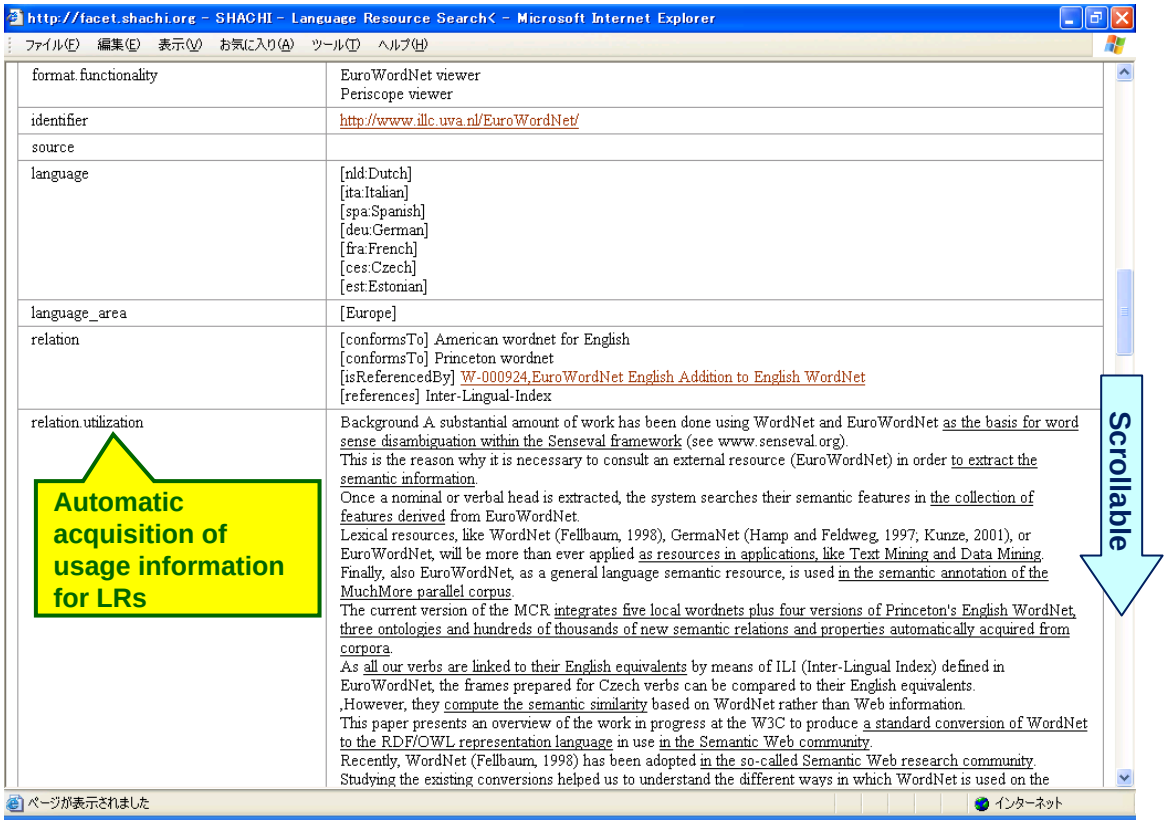

Figure1. A sample page of SHACHI catalog (ex. Euro WordNet)

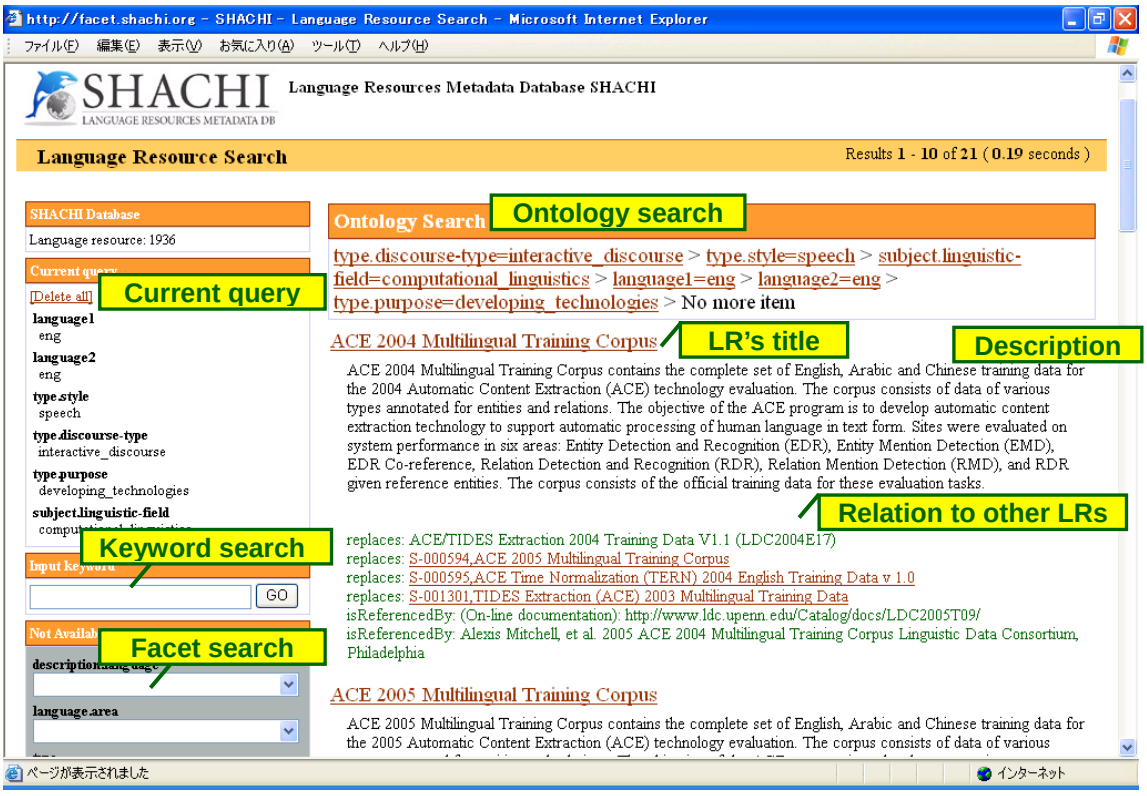

Figure 2. Catalogue search tool

5. Statistical data

It is possible to find the transition of characteristics that LRs having been released so far possessed by observing metadata of our collected LRs. This website provides the statistical information on the LR metadata database SHACHI on the spot. For example, Figure 3 is a graphic chart which shows an increasing tendency of numbers of LRs registered in SHACHI by language. The number of Chinese and Arabic LRs is notably increasing. Figure 4 shows an increasing tendency of written and spoken LRs. It can be observed that the number of released speech LRs has been increasing recently. Those kinds of information help us to promote strategic development of LRs. On SHACHI project, we are investigating what kind of LR is needed and how it is efficiently developed by utilizing them.

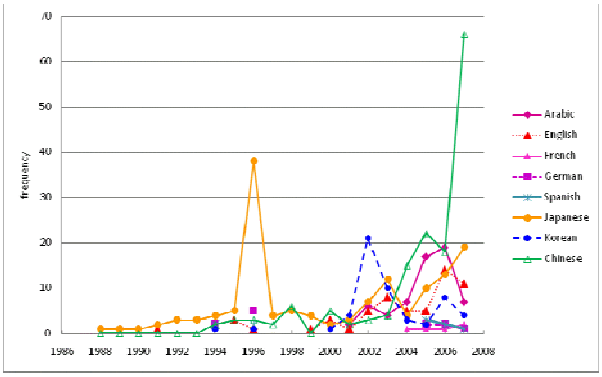

Figure 3. Increasing Tendency by language
（Indicated languages are those which have more than 15 LRs registered in SHACHI.）

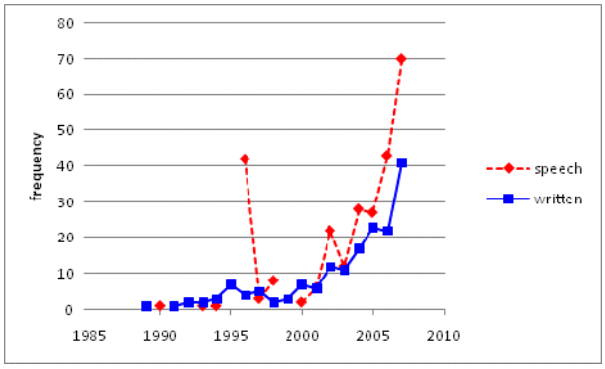

Figure 4. Increasing tendency of written and spoken LRs

6. Conclusion

In this research, we reported on the design of SHACHI, a metadata database of LRs now being developed, the expansion and construction of metadata for it, and an actualization of a search function we have developed. At present, it contains approximately 2,000 pieces of meta information on LRs such as corpora, dictionaries and thesauruses, each of which is fed into the database through semi-automatic means, forming a large scale LRs metadata archive. One of SHACHI's characteristic features is that with a collection of tag sets and format samples given to LRs it has a desirable design for an efficient development of LRs including the standardization of tags and organic combination of LRs. This collection of meta information has been reinforced by manually entering even more detailed meta information into the bottom-level category of the meta element set. We are currently attempting to measure the affinity between LRs and to systematically store and classify LRs in the world (construction of LRs ontology). From now on the SHACHI project is intended to promote cooperation among other LRs consortia abroad as well as in Japan and to take the initiative in contributing to the development of LRs in Asia.

Acknowledgements

The construction of this database presented in this paper was done in collaboration with Professor Kiyoaki Shirai at the Japan Advanced Institute of Science and Technology, Sachiko Waki and Cristobal Viveros as a translator, Miho Ohnishi at the Graduate School of Literature in Nagoya University, Takahiro Ono and Kenji Sugiki at Matsubara Group of the Graduate School of Information Science in Nagoya University and all the registration staff of Anchor Co., Inc. We would like to thank them for their generous support.

References

Hayashi, Y., Declerck, T., Buitelaar, P., Monachini, M. (2008) *Ontologies for a Global Language Infrastructure*, ICGL, pp.105-112.

Hughes, B., and Kamat, A. (2005). *A Metadata Search Engine for Digital Language Archives*, D-Lib Magazine, Volume 11 Number 2, http://www.dlib.org/dlib/february05/hughes/02hughes.html.

Ishida, T., Nadamoto, A., Murakami, Y., Inaba, R., Shigenobu, T., Matsubara, S., Hattori, H., Kubota, Y., Nakaguchi, T., and Tsunokawa, E. (2008). *A Non-Profit Operation Model for the Language Grid*, ICGL, pp.114--121.

Kozawa, S., Tohyama, H., Uchimoto, K., Matsubara, S., and Isahara, H. (2008). *Automatic Acquisition of Usage Information for Language Resources*, In Proceedings of the 6th edition of the Language Resources and Evaluation Conference, in print.

Tohyama, H., Kozawa, S., Uchimoto, K., Matsubara, S., and Isahara, H. (2008). *SHACHI: A Large Scale Metadata Database of Language Resources*. In Proceedings of the 1st International Conference on Global Interoperability for Language resources, pp. 205--212.