SHACHI: A Large Scale Metadata Database of Language Resources

Hitomi Tohyama†, Shunsuke Kozawa†, Kiyotaka Uchimoto††,
Shigeki Matsubara† and Hitoshi Isahara††
†Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
{hitomi,kozawa,matubara}@el.itc.nagoya-u.ac.jp,
††National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{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 in Asia and Western countries, for the purpose of effectively combining language resources. The purpose of this project is to investigate languages, tag sets, and formats compiled in language resources throughout the world, to systematically store language resource metadata, to create a search function for this information, and to ultimately utilize all this for a more efficient development of language resources. This metadata database contains more than 1,700 compiled language resources such as corpora, dictionaries, thesauruses and lexicons, forming a large scale metadata of language resources archive. Its metadata, an extended version of OLACmetadataSet 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.

1 Introduction

The importance of the construction of language resources such as corpora, dictionaries, thesauruses, etc., has been widely recognized, and has boomed for years throughout the world in its aim of encouraging research and development in the main media of spoken and written languages. Among the organizations willing to store and distribute language resources, there exist some consortia fulfilling their function such as LDC (Linguistic Data Consortium), ELRA (European Language Resources Association), OLAC (Open Language Archives Community) and Chinese-LDC (Chinese Linguistic Data Consortium) in Western countries and China, and GSK (Gengo Shigen Kyokai; Language Resource Association) which does so mainly in Japan. However, those released language resources are scarcely connected with each other not only because of the difference between written languages and spoken languages but also because of the difference between languages such as Japanese, English, Chinese, etc. Moreover, since language information tags given to those language resources and their data formats are multifarious, each language resource is operated individually.

As language resource development generally entails enormous cost, it is highly desirable that the research efficiency be enhanced by combining those existing language resources and systematically developing them altogether. For the purpose of fully integrating their language resources, NICT and Nagoya University have been constructing a large scale database named SHACHI as their joint project by collecting detailed meta information on language resources in Western and Asian countries.

This research project aims to extensively collect, and systematically store, metadata such as tag sets, formats, and language resource recordings existing at home and abroad. Meanwhile, we have already developed a facet search function of language resources using meta information, and are performing the experiment of widely providing this meta information on the stored language resources to those from professional researchers to common users. This paper outlines our language resource database, named SHACHI, in the development stage.

The structure of this paper is as follows: first we will outline the purpose and design of SHACHI in the second chapter. Next we will describe the collection of metadata in the third chapter, and the database structure, as well as the fundamental statistics, in the fourth chapter. Finally we will explain our future task in the fifth chapter.

2 Design of the Metadata Database

2.1 Purpose of the Construction of SHACHI

The purpose of the construction of SHACHI is fourfold.

- To investigate the actual conditions of tags and format types of language resources existing at home and abroad.
- To systematically obtain and store metadata of international language resources according to the information obtained from the above-mentioned. (This leads to the construction of language resource ontology.) (Hayashi, 2007)
- To conduct an investigation into the organic combination of language resources. (This leads to the strategic development of language resources.)
- To promote the distribution of language resources.

Some 2,000 resources of metadata have already been collected in the database so far, and by December 2008 they will be further enlarged by 3,000. 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.

Additionally, this database is obviously different from those of other language resource consortia since all of our detailed metadata are inputted manually. The database is notably characterized by the attempt to make an ontological construction of language resources throughout the world, as the affinity of language resource types and that of their tag sets are analyzed by applying natural language processing techniques to those detailed metadata. It seems certain that the realization of its ontological construction will contribute to a cutback in research and development costs, and to establishing a language resource infrastructure which meets present-day needs as an on-demand service. (NICT, 2007).

2.2 Design for Collecting Metadata

Among organizations willing to store and distribute language resources, there exist some consortia fulfilling their function such as LDC, ELRA, OLAC and Chinese-LDC in Western countries and China, and mainly GSK in Japan.

As for websites, there are two attempting to systematically amass metadata of language resources and promote their distribution, such as Language Technology World (LTW 2007) owned by DFKI (Deutsches Forschungszentrum for Kunstiliche Intelligencz) and a page owned by OLAC (OLAC, 2007).

To return the benefit of developed information processing technologies to society, it is highly desirable that the research be done in mutual cooperation among various language resource consortia and be enhanced by mutually exchanging information. SHACHI will make this possible as its metadata enables us to collect more detailed meta information in accordance with the OLAC metadata set by extending it. OLAC is creating a worldwide virtual library of language resources by developing consensus on the best current practice for the digital archiving of language resources, and by developing a network of interoperating repositories and services. OLAC metadata is based on the complete set of Dublin Core metadata set but a part of which was extended. The metadata set of SHACHI is described in detail in the following third chapter.

Table 1. SHACHI metadata set

	DC Qualifiers		Qualifiers for more precise description of the resources	
	LEVEL 1	LEVEL 2		
	DCMES Elements	DC Element Refinements	OLAC Extensions	SHACHI Extensions
1	title	alternative		
2	creator			
3	subject		linguistic-field (29) language (OLAC-Language extension)	mono-multi-lingual (2) monolingual multilingual resource-subject (3) dictionary thesaurus lexicon
4	description			price
5	publisher			
6	contributor		role(24)	mother-tongue (2) native non-native intonation (2) standard_dialect dialect level (2) age (3) gender (3)
7	date	created issued		
8	type		discourse-type (10) linguistic-type (3)	purpose(4) lexicography analysis developing_technologies education style (2) speech written form (2) fixed unfixed sentence(3) short long mixed has-annotation (2) annotated plain annotation sample
9	format	extent medium		encoding markup functionality
10	identifier			
11	source			
12	language		language (OLAC-Language extension)	
13	relation	DC relation refinements (13)		utilization
14	coverage	temporal		
15	rights			

Table 2. Description of SHACHI extensions

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
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_technologies	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.

3 Collecting Metadata

3.1 Extensions of Metadata Element

The metadata set of this language resource database follows 15 kinds of elements of Dublin Core (title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, rights), which is an extended version of OLAC metadata set, Table 1 shows the SHACHI metadata set. The shaded boxes "SHACHI Extension" are the elements added to those of Dublin Core. As for the definition of extended metadata elements, Table 2 is provided for showing it with the proviso that only one of the extended elements named "Utilization" is adopted on trial. This element is to provide information on the use of language resources which is semi-automatically retrieved from scholarly papers (Kozawa, 2007). Those items were added to this metadata, because it seemed that they gave us clues for finding the attributes of a corpus in collecting metadata of language resources such as corpora. Furthermore, we encourage our registrants to register as minutely as possible the information corresponding to each metadata element by themselves instead of automatically. In consequence, the information sometimes contains a lot of key words and somewhat longer sentences, which however enlarges the data subject to the key word search. We believe that its information will provide important clues to measure the affinity (similarity) of each language resource through language processing technologies in the future. The development of this project must contribute to the construction of an international language resource ontology.

As for the meta elements of language attributes, 160 values are currently introduced in our metadata set, which conforms to ISO 639 (ISO, 2007). The way to describe the date and the time conforms to ISO 8601. Moreover, this database will be intended to conform to ISO TC46/SC4, the International Standard for Information and Documentation & Technical Interoperability.

3.2 Policy for Collecting Metadata

The language resources which SHACHI stores should satisfy the following conditions:

1. The resources should be stored in a digital format.
2. The resources should be one of the followings: corpus, dictionary, thesaurus, or lexicon. (Numeric data are not considered to be the subject of collection for this database.)
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.

In addition to the above conditions, we are primarily collecting metadata of language resources which contain a large volume of data, are well known to the public, and are considered to be important for improving information processing technology. As expected we will call for participation in our project from research laboratories and institutes developing languages resources, by registering them in our database, as soon as preparations for publishing RSS feeds of our data and our input format of metadata standards are completed. Its registration page is also to be open to the public but is not yet on view.

- **Preferred Collection of Highly Recognized Language**

  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. Then, a search for language resources using key words such as “corpus”, “dictionary”, “thesaurus”, “translation” or “multilingual”, or key phrases such as “corpus so-and-so” (e.g. such as ‘PennTreeBank’), was done and as a result those which ranked highly and fitted our key words were retrieved. Table 3 shows the result. According to this research, WordNet was found to be the most widely distributed language resource used as a thesaurus. WordNet was originally compiled in certain European languages, and is now also being compiled in Chinese, Korean and other Asian languages. On the other hand, it was found that WordNet in Japanese has not been on the Internet yet. Research done through the Internet such as that described here is considered to be important in identifying the types of language resources to be developed in the future.

- **Collecting the Metadata from Major Organizations**

  It is important to gather language resource metadata from all over the world to perform a study about language resources, their promotion of distribution, and their strategic development. SHACHI not only covers metadata from major language resource consortia in Japan, Western countries, and China, but also conducts semi-automatic ways of registering detailed metadata in accordance with SHACHI metadata sets. Table 4 shows the list of major language resource consortia which this database covers.

4 Construction of SHACHI

SHACH is composed of language resource catalogs, a list of all language resource catalogs, catalogue search tool by which users can retrieve the information from all language resources stored in SHACHI from all angles, and the statistical information of the metadata of the language resources of SHACHI.

Table 3. The results of investigation on highly recognized and frequently used language resources

Monolingual Dictionary	Multilingual Dictionary	Parallel Corpus	Thesaurus
Webster's Revised Unabridged Dictionary (1913)	The EDICT Dictionary File	Aligned Hansards corpus	WordNet
Oxford Advanced Learner's Dictionary	WebLSD	European Parliament Proceedings Parallel Corpus 1996-2003	EuroWordNet
LONGMAN Dictionary of Contemporary English	Oxford-Hachette French Dictionary	OPUS - an open source parallel corpus	Hindi WordNet
Wiktionary	Collins ROBERT French Dictionary	UN Parallel Text	Roget's
Collins COBUILD Advanced Learner's Dictionary	Equine Multilingual Dictionary	Hong Kong Parallel Text	MeSH
The Swedish PAROLE Lexicon	The Papillon Project	COMPARA	Global WordNet
American Heritage Dictionary of the English Language	CJK Lexical Resources	The English-Norwegian Parallel Corpus	UMLS Metathesaurus
EDR Electronic Dictionary Technical Guide	Multilingual Dictionary of Proper Nouns CJKE-DPN	CRATER Multilingual Aligned Annotated Corpus	Merriam-Webster's Collegiate Thesaurus
Le Petit Robert French monolingual dictionary	Multilingual Glossary of technical and popular medical terms in nine European Languages	The JRC-Acquis Multilingual Parallel Corpus	ERIC Thesaurus
SANSEIDO Daijirin	*EIGIRO*	Polyglot Bible	Art & Architecture Thesuarus
A Japanese Lexicon (Japanese link)	EDR Bilingual Dictionary	NICT JLE Corpus	The European multilingual thesaurus on health promotion in 12 languages

Table 4. List of major language resource consortia which SHACHI covers

Consortium	ULR
Asian Language Resource Catalogue	http://nlp.kuee.kyoto-u.ac.jp/
ChineseLDC (ChineseLinguistic Data Consortium)	http://www.chineseldc.org/
ELRA(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 on CD-ROM	http://icame.uib.no/newcd.htm
LDC(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
SiTEC (Speech Information Technology & Industry	http://www.sitec.or.kr/English/
Speech Resources Consortium	http://research.nii.ac.jp/src/links/

The input format of metadata is, of course, not open to the public, but will be released in the future for use by research organizations and language resource consortia scattered all over the world. Because of this, we intend to provide outfits and interfaces which will allow researchers to register their language resource metadata without restraint, and which will enrich the quality of SHACHI by extending the contents of its database.

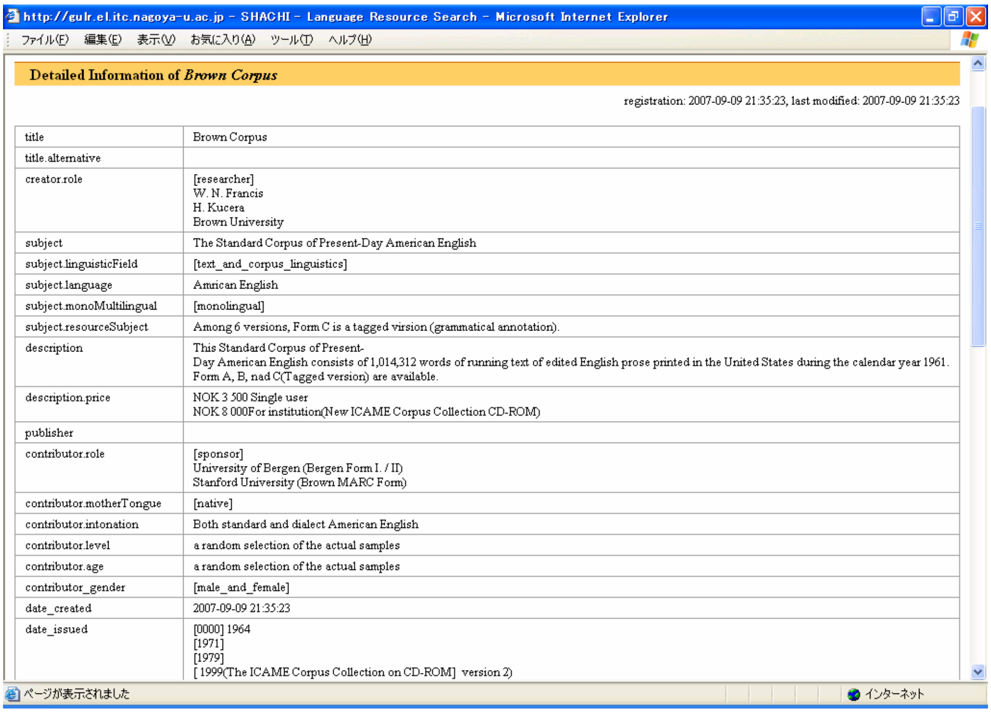
4.1 Outline of Catalogue

This catalogue allows SHACHI users to obtain outlines of the language resources at a glance without referring to the linked official websites (See Figure 1).

SHACHI catalogue allows SHACHI users to grasp outlines of the language resources at a glance without referring to the linked official websites.

4.2 Catalogue Search Tool

A catalogue search tool has been developed for the purpose of allowing users who visit the SHACHI site to find a language resource that meets their needs. This search tool is loaded with a key word search function as well as a facet search function. Figure 2 shows an image screen of the search tool. As for the key word search function, it allows users to input key words as they want, and to search all words stored in the SHACHI metadata archive. On the other hand, the facet search function is equipped with choices of 15 kinds of metadata elements selected from 25 kinds that SHACHI stores. This function helps users to obtain the desired language resource by selecting, in order, an element which is closest to their needs and then allows them to narrow down their choices. SHACHI especially brings its ability into full play for users who have no idea of adequate key words or have only a vague idea of how to find their desired information, since it can provide a large scale database with more detailed metadata through its characteristic search function (Bontcheva, 2007). On an actual image screen of the search results, titles of language resource candidates and a list of their descriptions are displayed, and at the bottom the relationships among other language resources are indicated (See Figure 2). We intend to conceive a method which enables relationships among language resource candidates


http://gulr.el.itc.nagoya-u.ac.jp - SHACHI - Language Resource Search - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
ページが表示されました
インターネット

Detailed Information of *Brown Corpus*

registration: 2007-09-09 21:35:23, last modified: 2007-09-09 21:35:23

title	Brown Corpus
title.alternative	
creator.role	[researcher] W. N. Francis H. Kucera Brown University
subject	The Standard Corpus of Present-Day American English
subject.linguisticField	[text_and_corpus_linguistics]
subject.language	Amrican English
subject.monoMultilingual	[monolingual]
subject.resourceSubject	Among 6 versions, Form C is a tagged virsion (grammatical annotation).
description	This Standard Corpus of Present- Day American English consists of 1,014,312 words of running text of edited English prose printed in the United States during the calendar year 1961. Form A, B, nad C(Tagged version) are available.
description.price	NOK 3 500 Single user NOK 8 000For institution(New ICAME Corpus Collection CD-ROM)
publisher	
contributor.role	[sponsor] University of Bergen (Bergen Form I. / II) Stanford University (Brown MARC Form)
contributor.motherTongue	[native]
contributor.intonation	Both standard and dialect American English
contributor.level	a random selection of the actual samples
contributor.age	a random selection of the actual samples
contributor_gender	[male_and_female]
date_created	2007-09-09 21:35:23
date_issued	[0000] 1964 [1971] [1979] [1999(The ICAME Corpus Collection on CD-ROM] version 2)

Figure 1. A sample page of SHACHI catalog (A part of catalogue; ex Brown Corpus)

shown in the search result to be visualized and measured.

4.3 Statistical Data

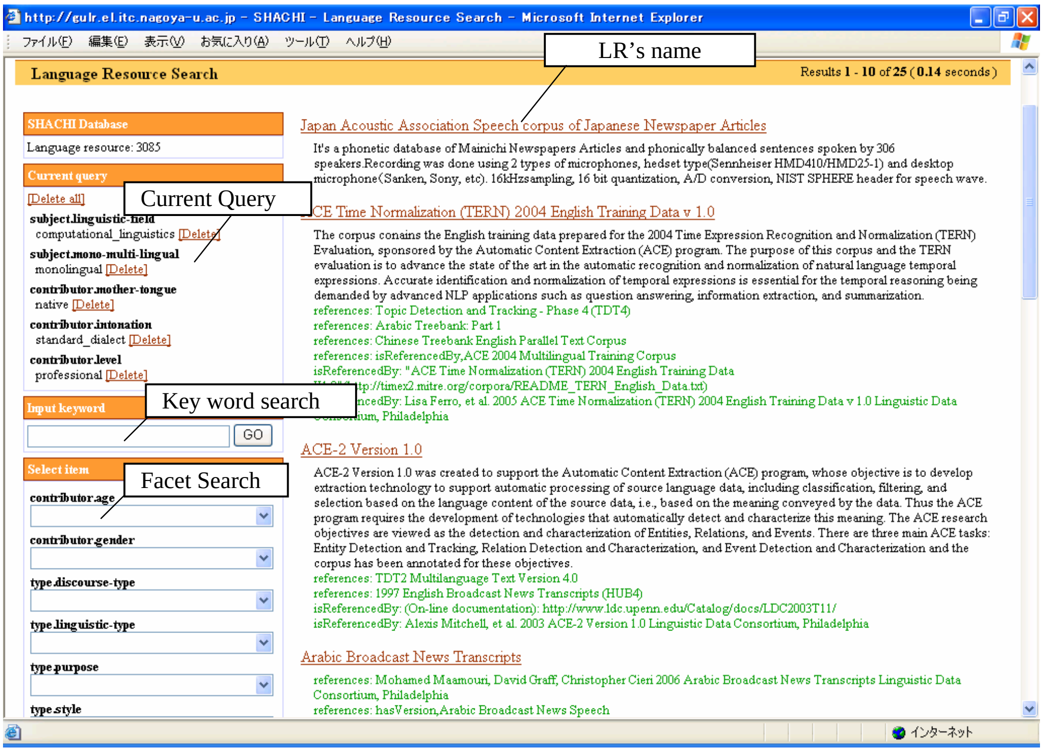
By observing metadata of the collected language resources, it is possible to find the transition of characteristics possessed by language resources that have been released so far. This website provides on the spot statistical information derived from the language resource metadata database SHACHI. It gives us the information, for example, that Arabian, Korean and Chinese language resources have been increasing recently, and that the prices of language resources have been rising. Thus, the SHACHI project investigates what kinds of language resources are needed and how they can be efficiently developed.

5 Future Work

These days new language resources are being created one after the other, and those resources are also modified on a daily basis. In order to grasp the current conditions of those resources, collecting and updating metadata at regular intervals is inevitably required. In addition, it is indispensable for us to collaborate with other research institutions such as GSK in order to collect language resource metadata efficiently and to develop the half-automated way of collecting and updating information via Web conversation (an online communication tool). On the other hand, we will conceive of an automatic information retrieval method by which we will obtain information on the situation of language resource diffusion, and various ways to utilize this information collected through the results of web searches and quoted information from papers.

6. Conclusion

The National Institute of Information and Communications Technology and Nagoya University have been collecting meta information extensively in order to construct a new type of language resource which uses metadata in their research through an organic combination of language resources. In this paper, the design of SHACHI (a metadata database of language resources now being developed), the expansion and construction of metadata for it, and the actualization of a search function we have developed were reported.



http://gulr.el.itc.nagoya-u.ac.jp - SHACHI - Language Resource Search - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
LR’s name
Current Query
Key word search
Facet Search
GO
インターネット

Language Resource Search

Results **1** - **10** of **25** (**0.14** seconds)

SHACHI Database

Language resource: 3085

Current query

[Delete all]

subject.linguistic-field
computational_linguistics [Delete]

subject.mono-multi-lingual
monolingual [Delete]

contributor.mother-tongue
native [Delete]

contributor.intonation
standard_dialect [Delete]

contributor.level
professional [Delete]

Input keyword

Select item

contributor.age

contributor.gender

type.discourse-type

type.linguistic-type

type.purpose

type.style

Japan Acoustic Association Speech corpus of Japanese Newspaper Articles

It's a phonetic database of Mainichi Newspapers Articles and phonically balanced sentences spoken by 306 speakers.Recording was done using 2 types of microphones, hedset type(Sennheiser HMD410/HMD25-1) and desktop microphone(Sanken, Sony, etc). 16kHzsampling, 16 bit quantization, A/D conversion, NIST SPHERE header for speech wave.

ACE Time Normalization (TERN) 2004 English Training Data v 1.0

The corpus conains the English training data prepared for the 2004 Time Expression Recognition and Normalization (TERN) Evaluation, sponsored by the Automatic Content Extraction (ACE) program. The purpose of this corpus and the TERN evaluation is to advance the state of the art in the automatic recognition and normalization of natural language temporal expressions. Accurate identification and normalization of temporal expressions is essential for the temporal reasoning being demanded by advanced NLP applications such as question answering, information extraction, and summarization.
references: Topic Detection and Tracking - Phase 4 (TDT4)
references: Arabic Treebank: Part 1
references: Chinese Treebank English Parallel Text Corpus
references: isReferencedBy,ACE 2004 Multilingual Training Corpus
isReferencedBy: "ACE Time Normalization (TERN) 2004 English Training Data
tp://timex2.mitre.org/corpora/README_TERN_English_Data.txt)
ncedBy: Lisa Ferro, et al. 2005 ACE Time Normalization (TERN) 2004 English Training Data v 1.0 Linguistic Data
ium, Philadelphia

ACE-2 Version 1.0

ACE-2 Version 1.0 was created to support the Automatic Content Extraction (ACE) program, whose objective is to develop extraction technology to support automatic processing of source language data, including classification, filtering, and selection based on the language content of the source data, i.e., based on the meaning conveyed by the data. Thus the ACE program requires the development of technologies that automatically detect and characterize this meaning. The ACE research objectives are viewed as the detection and characterization of Entities, Relations, and Events. There are three main ACE tasks: Entity Detection and Tracking, Relation Detection and Characterization, and Event Detection and Characterization and the corpus has been annotated for these objectives.
references: TDT2 Multilanguage Text Version 4.0
references: 1997 English Broadcast News Transcripts (HUB4)
isReferencedBy: (On-line documentation): http://www.ldc.upenn.edu/Catalog/docs/LDC2003T11/
isReferencedBy: Alexis Mitchell, et al. 2003 ACE-2 Version 1.0 Linguistic Data Consortium, Philadelphia

Arabic Broadcast News Transcripts

references: Mohamed Maamouri, David Graff, Christopher Cieri 2006 Arabic Broadcast News Transcripts Linguistic Data Consortium, Philadelphia
references: hasVersion,Arabic Broadcast News Speech

Figure 2. Catalogue search tool

SHACHI contains extensive data obtained from highly recognized language resources searched for on the Internet, and also covers meta information provided by language resources consortia such as ELRA, OLAC and LDC. Furthermore, this continuing work to register more detailed metadata is making SHACHI an even larger scaled database. At present, it contains approximately 1,700 pieces of meta information on language resources, forming the world largest language resource metadata archive. One of SHACHI’s characteristic features is that the manner in which it collects tag sets and format samples given to language resources, has a desirable design for the strategic development of language resources including the standardization of tags and the efficient development of language resources. This collection of meta information has been further reinforced by the manual entering of even more detailed meta information into the bottom-level category of the meta element set. We have already measured the affinity between language resources and have systematically stored and classified language resources throughout the world. We believe this will lead to the construction of a language resource ontology.

Acknowledgements

We would like to thank Sachiko Waki, Miho Ohnishi of Nagoya University, Kenji Sugiki and Takahiro Ono of Matsubara’s SLP Group at Nagoya University and all the staff of ANCHOR Co., Inc. for their full cooperation in constructing the database.

References

Hayashi, Y. 2007. Conceptual Framework of an Upper Ontology for Describing Linguistic Services, Proceedings of IWIC2007, pp.246-260.

ISO official site, 2007.
http://www.iso.org/error/sitedown.html

Kozawa, S. Toyama, H. Uchimoto, K. Matsubara, S. 2007. Automatic Acquisition of Expressions Representing Purposes and Methods of Using A Language Resource from Academic Articles, Proceedings of Workshop on Informatics 2007, pp.65-70, (in Japanese).

LTW (Language Technology World). 2007. official site, http://www.dfki.de/lt//publications_show.php?id=148

NICT Language Grid project official site, 2007 . official site, http://langrid.nict.go.jp/indexj.htm.

OLAC (Open Language Archives Community), 2007. official site, http://www.lt-world.org/