# Storage of Usage Information on Research Resource Metadata Database: Corpus Construction Using Academic Articles

**Shunsuke Kozawa**[1], **Hitomi Tohyama**[2],
**Kiyotaka Uchimoto**[3], and **Shigeki Matsubara**[4], Non-members

## ABSTRACT

Recently, language resources have become indispensable for linguistic researches. However, existing language resources are seldom fully utilized because their variety of usage is not well known, indicating that their intrinsic value is not recognized very well either. Regarding this issue, lists of usage information might improve language resource searches and lead to their efficient use. In this research, therefore, we collect a list of usage information for each language resource from academic articles to promote the efficient utilization of language resources. This paper describes the construction of a text corpus annotated with usage information (UI corpus). In particular, we automatically extract sentences containing language resource names from academic articles. Then, the extracted sentences are annotated with usage information by two annotators in a cascaded manner. We will show that the UI corpus contributes to efficient language resource searches, by combining the UI corpus with a metadata database of language resources and comparing the number of language resources retrieved with and without the UI corpus.

**Keywords**: Research Resource, Language Resource, Usage Information, Academic Articles

## 1. INTRODUCTION

In recent years, such many research resources as samples, materials, software and databases have been developed and published by researchers. However, existing research resources are not fully utilized. In this paper, we focus on language resources. Such language resources as corpora and dictionaries are being widely used for research in the fields of linguistics, natural language processing, and spoken language processing, reflecting the recognition that objectively analyzing linguistic behavior based on actual examples is important. Therefore, since the importance of language resources is widely recognized, they have been constructed as a research infrastructure and are becoming indispensable for linguistic research.

The development of language resources generally entails enormous cost and effort. To encourage efficient development of linguistic technologies and language resources, it is highly desirable that existing language resources be not only used for their developer's purposes but also widely shared and used in many fields. There exist the organizations willing to store and distribute language resources such as Linguistic Data Consortium (LDC)[1] , European Language Resources Association (ELRA)[2] , Common Language Resources and Technologies Infrastructure (CLARIN)[3] [1], ACL Data and Code Repository[4] , and Open Language Archives Community (OLAC)[5] [2]. Moreover, in order to enhance efficient use of language resources, metadata search services for language resources [3, 4] and web services for language resources [5, 6, 7, 8, 9] have become available. However, since language information tags given to those language resources and their data formats are multifarious, each language resource is operated individually and existing language resources are scarcely connected with each other. Therefore, it is difficult for researchers and users to understand the relationships among language resources and find language resources which might be useful for their researches.

In this issue, the ontologies or categories of language resources have been constructed using the metadata of language resources [10, 11, 12, 13]. However, it has not been enough for researchers and users to efficiently find and use language resources suitable for their own purposes so far. If there exists a system which could give us a list of language resources that can be answers to questions such as "Which language resources can be used for developing a syntactic parser?" and "Which language resources can be used for developing a Chinese-English machine translation system?," it would help users efficiently find appro-

Manuscript received on August 11, 2011 ; revised on September 15, 2011.

[1,2,4] The authors are with Graduate School of Information Science, Nagoya University Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan, Email: kozawa@el.itc.nagoya-u.ac.jp, hitomi@el.itc.nagoya-u.ac.jp and matubara@nagoya-u.jp

[3] The author is with National Institute of Information and Communications Technology 3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan , Email: uchimoto@nict.go.jp

[1]LDC: http://www.ldc.upenn.edu/
[2]ELRA: http://www.elra.info/
[3]CLARIN: http://www.ilsp.gr/clarin_eng.html
[4]ACL Wiki: http://www.aclweb.org/aclwiki/
[5]OLAC: http://language-archives.org/

priate language resources.

A few projects have collected the information satisfying these demands, that is, information about what language resources are used for (i.e. "usage information"). In Language Technology World (LT World)[6] , for technology terms, e.g., "Information Extraction" and "Machine Translation", extracted from academic articles, their area experts provide the relevant projects and resources [12]. In LREC Map[7] , the conference organizers have had the authors post information on language resources used or developed in their papers during the submission of their papers to international conferences in the field of computational linguistics such as the conference on language resources and evaluation (LREC), the conference on computational linguistics (COLING) and the conference on empirical methods in natural language processing (EMNLP) [14]. In SHACHI[8] , the steering body manually collected usage information described on official web sites of individual language resources [15].

In this research, we focus on SHACHI since SHACHI manages and provides usage information for language resources in an integrated fashion. SHACHI is the metadata database of language resources and contains metadata on approximately 2,400 language resources. SHACHI metadata set is an extended version of OLAC metadata set which conforms to Dublin Core[9] [16] metadata element set. Usage information manually collected from official web sites has been described in the 'type.purpose' which is one of the extended metadata. However, the number of lists of usage information registered in SHACHI is only about 900 language resources since the usage information is not usually described on the official site while it is often described in academic articles. For instance, the following sentence found in the proceedings of ACL2006 shows that Roget's Thesaurus is useful for word sense disambiguation, although usage information is not announced on the web page of Roget's Thesaurus[10].

- He also employed *Roget's Thesaurus* in 100 words of window to implement **WSD**.

Therefore, we could more easily find language resources suitable for our own purposes by collecting lists of usage information for language resources from academic articles and integrating them with metadata contained in SHACHI. Although the method for automatically extracting the lists was proposed [17], the variation of the extracted usage information was limited since their extraction rules were based on the analysis of small lists of usage information. This issue would be addressed by collecting large lists of usage information and then analyzing them to expand the extraction rules. Therefore, in this paper, we describe the construction of a text corpus annotated with usage information (UI corpus).

This paper is organized as follows: In section 2, we introduce the design of UI corpus. Then, we explain the construction of the UI corpus by extracting sentences containing language resource names from academic articles and annotating them with usage information in section 3. In section 4, we evaluate the UI corpus by investigating whether annotated usage information is correct. In sections 5 and 6, we provide statistics of the UI corpus and analytical results of usage information contained in the UI corpus. We show that the UI corpus contributes to efficient language resource searches by combining the UI corpus with a metadata database of language resources and comparing the number of language resources retrieved with and without the UI corpus in section 7. Finally, in section 8, we describe the summary of this paper and the future work.

## 2. DESIGN OF THE UI CORPUS

### 2.1 Data Collection

It is unrealistic to collect all sentences in academic articles and annotate them because only a small number of sentences in an article include usage information for language resources. In this issue, Kozawa et al. reported that most of the instances of usage information for language resources are found in the sentences containing language resource names [17]. Therefore, in this research, we collect sentences having language resource names from academic articles to build the UI corpus.

### 2.2 Annotation Policy

The collected sentences are annotated with the following information: (A), (B) and (C) are provided for each sentence. (D) and (E) are provided for word sequences. (A) through (D) are automatically provided when sentences have been collected from academic articles.

A) sentence ID

B) the title of the proceeding

C) article ID

D) language resource name
Word sequences matched with language resource names are annotated with LR tags as the following example:
- For comparison, <LR>Penn Treebank< /LR> contains over 2400 (much shorter) WSJ articles.

[6]LT World: http://www.lt-world.org/
[7]LREC Map: http://www.resourcebook.eu/LreMap/
[8]SHACHI: http://www2.shachi.org/
[9]http://dublingcore.org/
[10]http://poets.notredame.ac.jp/Roget/about.html

Note that the LR tags with which word sequences are automatically annotated need to be examined.

This is because homographs of the language resource names (ex. the names of projects and associations) are sometimes erroneously annotated as language resource names. For example, "Penn Treebank" is often used as a language resource name and it is sometimes used as a project name in a different context. Therefore, it is difficult to discriminate proper language resource names from others. Then, if inappropriate word sequences are annotated with LR tags, they are manually eliminated.

If word sequences matched with two or more language resources have a coordinate structure as the following example, "Chinese" and "English PropBanks" are annotated with LR tags. This is because we distinguish between two or more language resource names (e.g. Chinese and English PropBanks) and a language resource name which has a coordinate structure (e.g. Cobuild Concordance and Collocations Sampler)

- The functional tags for <LR>Chinese< /LR> and <LR>English PropBanks< /LR> are to a large extent similar.

E) usage information
Each word sequence matched with usage information for a certain language resource is annotated with UI tags. Word sequences are annotated with usage information by referring only to a given sentence without adjacent sentences in order to reduce the labor costs of annotators. In our research, we assume that usage information A for language resource X can be paraphrased as "X is used for A." The followings are examples of usage information for WordNet:

- We use <LR>WordNet< /LR> for <UI>lexical lookup< /UI>.
- <LR>WordNet< /LR><UI>specifies relationships among the meaning of word < /UI>.
- It uses the content of <LR>WordNet< /LR> to <UI>measure the similarity or relatedness between the senses of a target word and its surrounding words< /UI>

Note that since usage information indicates specific events, such vague expressions as "our proposed method" and "this purpose" are not our target. We also ignore the usage information expressions that represent updating, expansion, or modification of language resource X, as shown in the following example:

- We applied an automatic mapping from <LR>WordNet 1.6< /LR> to <LR>WordNet 1.7.1< /LR> synset lables.

## 3. CONSTRUCTION OF THE UI CORPUS

This section describes the method for constructing the UI corpus. Figure 1 shows the flow of the corpus construction. First, sentences containing language resource names are automatically extracted from academic articles. Then, the extracted sentences are annotated by two annotators in a cascaded manner.

### 3.1 Automatic Extraction of Sentences Containing Language Resource Names

First, we converted academic articles into plain texts using the Xpdf[11] . We used 2,971 articles which are contained in the proceedings of ACL (annual meeting of the association for computational linguistics) from 2000 to 2006, LREC2004 (international conference on language resources and evaluation) and LREC2006 because language resources are often used in the field of computational linguistics. Next, we extracted sentences containing the language resource names from the articles. As for the language resource names, approximately 2,400 language resources registered in SHACHI [15] were used. Out of 2,971 articles, 1,848 have sentences containing language resource names. Consequently, 10,959 sentences were extracted from 1,848 articles and word sequences matched with the language resource names are automatically annotated with LR tags.

***Table 1:*** *Size of the UI corpus*

| Item | Number |
|---|---|
| articles | 1533 |
| sentences | 8135 |
| LR tags | 10504 |
| sentences containing usage information | 1110 |
| UI tags | 1183 |

### 3.2 Annotation of Word Sequences with Usage Information

Two annotators were involved in the annotation. One of the annotators (Annotator 1) had an experience in collecting metadata in SHACHI, while the academic major of the others (Annotator 2) was computational linguistics. Since Annotator 1 was unfamiliar with computational linguistics, it was difficult for Annotator 1 to annotate usage information if a given sentence contains technical terms in the field of computational linguistics. Therefore, Annotator 1

[11] http://www.foolabs.com/xpdf/

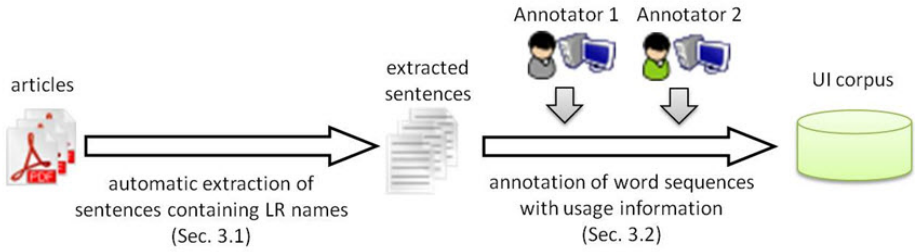


***Fig.1:*** *Flow of the UI corpus construction*

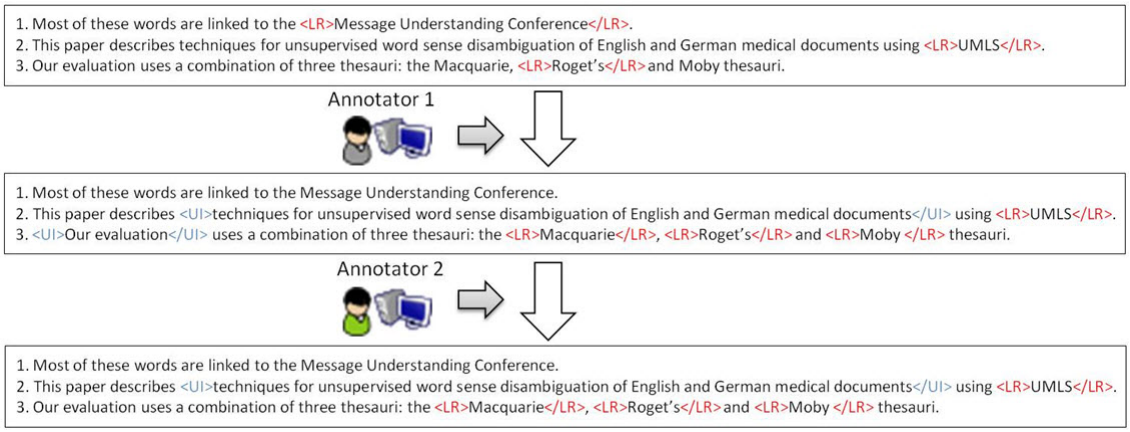


***Fig.2:*** *Flow of corpus annotation*

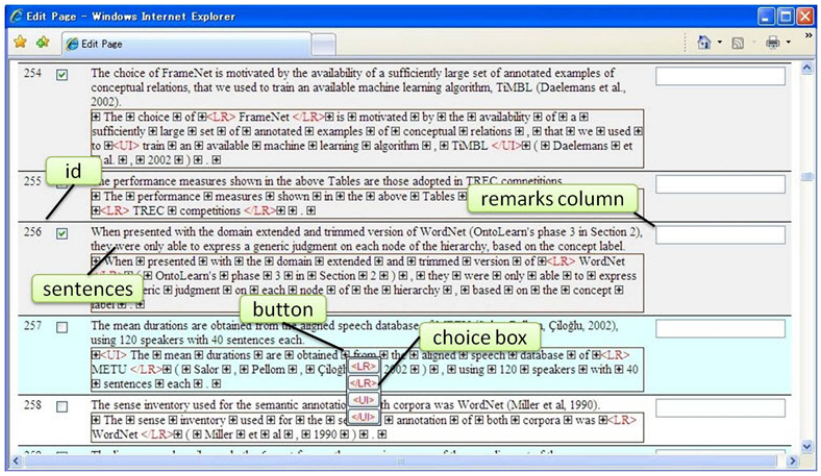


***Fig.3:*** *Web-based GUI for supporting annotation*

annotated sentences at first, and then Annotator 2 annotated the same sentences to correct the annotation errors produced by Annotator 1.

Examples of corpus annotation are shown in Figure 2. First, the following actions were done by Annotator 1:

- Annotating word sequences representing language resource names with LR tags
- Annotating word sequences representing the usage information for language resources with UI tags

Next, Annotator 2 judged whether the UI tags provided by Annotator 1 were correct and modified them if they were inappropriate.

For annotation, the annotators used the Web-based GUI as shown in Figure 3. They could make annotation by selecting an appropriate tag from the choice box appeared by clicking the button shown in Figure 3.

### 3.3 Corpus Size

The size of the annotated corpus is shown in Table 1. In the process of annotation, sentences which did not contain any language resource names were

removed from the corpus. Therefore, the corpus consists of a set of sentences containing language resource names extracted from academic articles.

## 4. EVALUATION OF THE UI CORPUS

To evaluate the validity of the UI corpus, we had another two annotators who are experts in computational linguistics annotate the sentences in the UI corpus with UI tags. Given 100 sentences randomly extracted from sentences annotated with UI tags and 100 sentences randomly extracted from sentences which were not annotated with UI tags, the two annotators annotated word sequences representing the usage information for language resources with UI tags in a cascaded manner as shown in Table 2. In this section, we call the annotators who constructed the UI corpus in section 3.2 and the

***Table 2:*** *The results of evaluation of the UI tags*

| Item | Number |
|---|---|
| agreement | 86 |
| disagreement | 46 |
| partial match | 9 |
| only Annotators A | 12 |
| only Annotators B | 25 |

***Table 3:*** *Frequently appearing language resources*

| Rank | language resource name | # of LR tags |
|---|---|---|
| 1 | WordNet | 2356 |
| 2 | Penn Treebank | 745 |
| 3 | FrameNet | 437 |
| 4 | British National Corpus | 433 |
| 5 | PropBank | 419 |
| 6 | SemCor | 209 |
| 7 | VerbNet | 191 |
| 8 | TREC Collection | 166 |
| 9 | MeSH | 147 |
| 10 | EuroWordNet | 144 |

annotators who annotated the sentences with UI tags in this section Annotators A and Annotators B, respectively.

The 200 sentences annotated by Annotators A have 107 UI tags, while those annotated by Annotators B have 120 UI tags. The agreement rate between Annotators A and B was 75.8% as shown in Table 2.

We investigated the UI tags which have disagreement between Annotators A and B to learn the causes of them. There were three types of the UI tags as follows:

- Partial match
  UI tags whose positions (the start tag <UI> and/or the end tag < /UI>) provided by Annotators A were different from those provided by Annotators B
- Only Annotators A
  UI tags provided by only Annotators A
- Only Annotators B
  UI tags provided by only Annotators B

The agreement rate was 83.7% if we considered the partial matching between Annotators A and B as agreement. Out of them, the UI tags provided by either Annotators A or B have a great influence on the validity and coverage of the UI tags. The number of them was 37 (16.3%), which is little. This indicates that the UI corpus has high quality although some of UI tags should be modified.

## 5. STATISTICS OF THE UI CORPUS

This section shows the frequently used language resources and the difference between the language resources tagged with UI tags in the UI corpus and the language resources whose usage information were registered in SHACHI by comparing the statistics of the UI corpus with those of SHACHI.

***Table 4:*** *Classes of usage information*

| | |
|---|---|
| Acoustic Modeling | Question Answering |
| Applications | Robust ASR |
| Asian Language Processing | Segmentation |
| Chunking | Semantics |
| Coreference | Speaker Recognition |
| Corpora | Speaker Segmentation |
| Dialect Recognition | Speech Analysis |
| Dialogue | Speech Coding |
| Discourse | Speech Enhancement |
| Generation | Speech Features |
| Grammars | Speech Perception |
| Information Extraction | Speech Processing |
| Information Retrieval | Speech Production |
| Language Acquisition | Speech Synthesis |
| Language Modeling | Speech Translation |
| Language Recognition | Spoken Language Processing |
| Large Vocabulary Speech Recognition | Spoken Language Resource |
| Lexical Semantics | Spoken Language Understanding |
| Linguistics | Statistical Machine Translation |
| Machine Translation | Statistical Parsing |
| Morphology | Summarization |
| Named Entity Recognition | Syntax |
| Parsing | Tagging |
| Phonetics | Text Categorization |
| Phonology | Word Sense Disambiguation |
| Prosody | |

We semi-automatically assigned each LR tag with a language resource id used in SHACHI and counted the number of language resources appearing in our corpus. Consequently, 882 language resources were found. Then, we investigated the breakdown of the LR tags. We found that the most frequently appearing language resource was WordNet (see Table 3). One of the reasons is that WordNet has been frequently used as a lexical database [18]. Another reason is that it has been translated into various languages by the initiative of Global WordNet Association[12].

We investigated language resources tagged with UI tags. Note that all 882 language resources appearing in the UI corpus do not have usage information since in the UI corpus, only their explanations or vague

[12]http://www.globalwordnet.org/

***Table 5:*** *Frequently appearing language resources*

| class | # of UI tags | # of articles | # of LRs | frequently used language resource |
|---|---|---|---|---|
| Lexical Semantics | 188 | 110 | 35 | WordNet(82), UMLS(4), EuroWordNet(4) |
| Word Sense Disambiguation | 182 | 88 | 42 | WordNet(61), SemCor(8), Roget's(5), Longman Dictionary(5) |
| Semantics | 147 | 81 | 38 | WordNet(27), FrameNet(18), PropBank(7) |
| Corpora | 130 | 66 | 51 | WordNet(28), Penn Treebank(4), VerbNet(4) |
| Information Extraction | 104 | 71 | 43 | WordNet(22), British National Corpus(6), ACE(5), GENIA(4) |
| Parsing | 66 | 44 | 30 | Penn Treebank(22), CCGbank(3) |
| Tagging | 59 | 40 | 29 | Penn Treebank(11), GENIA(4), CGN(4) |
| Question Answering | 48 | 26 | 13 | WordNet(15), TREC(5), Wikipedia(2), Extended WordNet(2) |
| Machine Translation | 44 | 32 | 31 | WordNet(7), Czech Treebank(2), British National Corpus(2) |
| Asian Language Processing | 36 | 30 | 24 | Mainichi newspaper(3), Sinica Treebank(2), EDR(2) |
| Grammars | 30 | 22 | 8 | Penn Treebank(17), British National Corpus(2) |
| Information Retrieval | 29 | 22 | 12 | WordNet(9), TREC(4), NTCIR(2), EuroWordNet(2) |
| Coreference | 29 | 18 | 15 | WordNet(4), British National Corpus(2), Reuters corpus(2) |
| Named Entity Recognition | 25 | 19 | 18 | MUC(4), GENIA(3), UMLS(2), WordNet(2), Reuters corpus(2) |
| Statistical Parsing | 25 | 15 | 7 | Penn Treebank(10), CCGbank(2) |
| Syntax | 18 | 15 | 10 | Penn Treebank(9) |
| Spoken Language Processing | 17 | 11 | 12 | Spoken Dutch Corpus(2), Penn Treebank(2) |
| Dialogue | 17 | 11 | 9 | ICSI Meeting corpus(2), British National Corpus(2) |
| Chunking | 16 | 9 | 8 | Penn Treebank(5) |
| Summarization | 14 | 11 | 10 | Mainichi newspaper(3), English Broadcast News corpus(2) |
| Generation | 14 | 9 | 9 | WordNet(3) |
| Text Categorization | 12 | 12 | 10 | Reuters-21578 corpus(3), WordNet(2), Gigaword corpus(2) |
| Morphology | 11 | 9 | 8 | Arabic Treebank(3), Chinese Penn Treebank(2) |
| Robust ASR | 10 | 8 | 9 | SpeechDat(1), LC-Star(1), ATIS corpus(1) |
| Language Modeling | 10 | 8 | 5 | British National Corpus(3), Yomiuri Newspapers(2) |
| Language Acquisition | 9 | 7 | 5 | WordNet(3) |
| Statistical Machine Translation | 9 | 6 | 8 | WordNet(2) |
| Prosody | 9 | 6 | 8 | Spoken Dutch Corpus(1), TDT(1), TTS evaluation database(1) |
| Segmentation | 9 | 4 | 8 | British National Corpus(1), Penn Treebank(1), ECI corpus(1) |
| Phonetics | 8 | 6 | 5 | Spoken Dutch Corpus(2), CELEX(2) |
| Linguistics | 8 | 5 | 5 | WordNet(2) |
| Phonology | 6 | 5 | 3 | CELEX(2), Penn Treebank(2) |
| Discourse | 4 | 3 | 3 | British National Corpus(2) |
| Large Vocabulary Speech Recognition | 3 | 3 | 3 | SpeechDat(1), Slovenian broadcast news speech database(1) |
| Applications | 3 | 3 | 3 | WordNet(1), Brown Corpus(1), Mainichi Daily News(1) |
| Speech Processing | 3 | 2 | 2 | Switchboard(1), RT corpus(1) |
| Speech Synthesis | 2 | 2 | 2 | LC-Star(1), METU Turkish Corpus(1) |
| Acoustic Modeling | 2 | 1 | 1 | NIST Corpus(1) |
| Speech Analysis | 1 | 1 | 1 | Czech National Corpus(1) |

usages (the underlined part) which do not indicate specific events were described for some language resources as follow:

- •<LR>BNSC< /LR> is a broadcast news corpus developed in the framework of the European-funded project Network of Data Centres (NetDC).
- • The third dataset used in our evaluation contains 25 meeting transcripts from the <LR>ICSI-MR corpus< /LR>.

Out of 882 language resources, 365 were tagged with UI tags. We investigated whether the usage information for the language resources was registered in SHACHI or not, and found that usage information for 305 language resources was not registered in SHACHI. This shows that usage information for language resources newly extracted from academic articles were almost double of that originally registered in SHACHI. We expect that more usage information could be extracted if we used more variety of academic articles and it would help users efficiently find and use language resources suitable for their own purposes by registering lists of usage information for finding more language resources than those obtained only with SHACHI.

## 6. ANALYSIS OF USAGE INFORMATION

This section shows the number of types of usage information and language resources used in various fields. Lists of usage information were analyzed to know how many types of usage information were collected. In this paper, we assumed that types of usage information for language resources can be represented by using session names appeared in international conferences in the fields of computational linguistics and spoken language processing. We extracted 51 classes, as shown in Table 4, which appeared twice or more as session names in the proceedings of ACL from 2000 to 2006 or ICSLP (international conference on spoken language processing) 2000, 2002, 2004 and 2006. Then, we manually classified lists of usage information into the classes by referring to the articles containing target usage information. Note that each list of usage information is classified into one or more classes.

***Table 6:*** *The results of evaluation of the UI tags*

| Rank | language resource name | # of classes |
|---|---|---|
| 1 | Penn Treebank | 24 |
| 2 | WordNet | 23 |
| 3 | British National Corpus | 21 |
| 4 | Reuter Corpus | 9 |
| 5 | EuroWordNet | 8 |
| | FrameNet | 8 |
| | UMLS | 8 |
| | Chinese Treebank | 8 |
| 9 | Spoken Dutch Corpus | 7 |
| | TREC Collection | 7 |
| | Brown Corpus | 7 |
| | TIGER Corpus | 7 |

Classification results are shown in Table 5. Columns 2 and 3 represent the number of UI tags classified into each class and the number of articles

containing the UI tags, respectively. The number of language resources tagged with UI tags is shown in column 4. In column 5, frequently used language resources in each field are represented and parenthetical figures denote the number of articles using the language resource. Large lists of usage information in the fields of "lexical semantics" and "word sense disambiguation" were collected since WordNet was frequently used. Out of 51 classes, 39 have one or more UI tags. This shows that the UI corpus contains a variety of usage information.

We investigated the number of classes to which UI tags were classified for each language resource to find language resources used in various fields. The results of the investigation are shown in Table 6. We found that Penn Treebank is most widely used language resource. In addition, British National Corpus has also been widely used although the frequency of British National Corpus is lower than those of WordNet and Penn Treebank.

## 7. CONTRIBUTION OF THE UI CORPUS

We carried out experiments of searching for language resources on the database using keywords to learn whether the number of retrieved language resources increases by using usage information in the UI corpus. As queries for language resource search, we used 40 keywords manually extracted from the "Topics of Interest" appearing in the paper submission page of ACL2008[13] since we assumed that researchers in the fields of computational linguistics searches for language resources suitable for their own purposes. We compared the number of language resources retrieved with and without usage information in the UI corpus. In the experiments, we used keywords as queries and got a list of language resources whose usage information registered in SHACHI or in the UI corpus contained the keywords.

The experimental results are shown in Table 7. The number of language resources retrieved using only usage information contained in SHACHI is shown in column 2 and the number of language resources retrieved using usage information contained in both SHACHI and the UI corpus is shown in column 3. The number of language resources retrieved using usage information in both SHACHI and the UI corpus increased for 15 keywords, which are attached with * in Table 7, compared to SHACHI. This indicates that lists of usage information in the UI corpus contribute to efficient language resource searches.

We are planning to train the model for extracting usage information for language resources by using our corpus to improve the performance of automatic usage information extraction and extract usage information from various articles. Then, we expect that more various language resources can be found.

[13]http://www.ling.ohio-state.edu/acl08/cfp.html

## 8. CONCLUSION

In this paper, we described how to construct the UI corpus to efficiently find and use appropriate language resources. First, we automatically extracted sentences containing language resource names from academic articles. Then, two annotators tagged the extracted sentences with usage information. We showed that the UI corpus contributes to efficient language resource searches by combining the UI corpus with a metadata database of language resources.

***Table 7:*** *Frequently appearing language resources*

| Keywords | SHACHI | SHACHI +UI corpus |
|---|---|---|
| *dialogue | 8 | 12 |
| embodied conversational agents | 0 | 0 |
| language-enhanced platforms | 0 | 0 |
| *information retrieval | 52 | 54 |
| text data mining | 0 | 0 |
| information extraction | 11 | 11 |
| *filtering | 0 | 2 |
| recommendation | 0 | 0 |
| *question answering | 0 | 3 |
| topic classification | 0 | 0 |
| text classification | 3 | 3 |
| sentiment analysis | 0 | 0 |
| attribute analysis | 0 | 0 |
| genre analysis | 0 | 0 |
| language generation | 1 | 1 |
| *summarization | 8 | 14 |
| *machine translation | 55 | 57 |
| language identification | 21 | 21 |
| multimodal processing | 0 | 0 |
| *speech recognition | 211 | 228 |
| speech generation | 1 | 1 |
| *speech synthesis | 32 | 34 |
| phonology | 0 | 0 |
| *POS tagging | 1 | 7 |
| *syntax | 0 | 7 |
| *parsing | 11 | 19 |
| grammar induction | 0 | 0 |
| mathematical linguistics | 0 | 0 |
| formal grammar | 0 | 0 |
| *semantics | 1 | 5 |
| textual entailment | 0 | 0 |
| *paraphrasing | 0 | 2 |
| *word sense disambiguation | 0 | 11 |
| *discourse | 10 | 14 |
| pragmatics | 0 | 0 |
| statistical and machine learning | 0 | 0 |
| language modeling | 25 | 25 |
| lexical acquisition | 0 | 0 |
| knowledge acquisition | 0 | 0 |
| development of language resources | 0 | 0 |

In the near future, we will provide a language resource search service to promote the efficient use of language resources by integrating usage information with a metadata database of language resources called SHACHI [15]. However, there exists the issue that the extended metadata are not normalized although the metadata which are the same as OLAC metadata are already normalized since SHACHI metadata set is an extended version of OLAC metadata set. We need to normalize the metadata set for language resources by consulting with other existing major language resource consortiums such as ELRA and LDC. We would like to suggest that the extended metadata be adopted as the normalized metadata to other language resource consortiums (some of the metadata have already been adopted in other language resource consortiums) since the extended metadata are designed for providing users and researchers with sufficient information for language resources.

## References


[1] T. Varadi, P. Wittenburg, S. Krauwer, M. Wynne and K. Koskenniemi, "CLARIN: Common Language Resources and Technology Infrastructure," in *Proceeding of 6th International Conference on Language Resources and Evaluation*, 2008.

[2] G. Simons and S. Brid, "The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources," *Literary and Linguistic Computing*, Vol. 18, pp. 117-128, 2003.

[3] P. Wittenburg, D. Broeder, F. Offenga and D. Willems, "Metadata Set and Tools for Multimedia/Multimodal Language Resources," in *Proceeding of 3rd International Conference on Language Resources and Evaluation*, 2002.

[4] B. Hudges and A. Kamat, "A Metadata Search Engine for Digital Language Resources," DLib Magazine, 11(2):6, 2005.

[5] A. Dalli and V. Tablan and K. Bontcheva and Y. Wilks and D. Broeder and H. Brugman and P. Wittenburg, "Web Services Architecture for Language Resources," in *Proceeding of 4th International Conference on Language Resources and Evaluation*, 2004.

[6] C. Biemann, S. Bordag, U. Quasthoff and C. Wolff, "Web Services for Language Resources and Language Technology Applications," in *Proceeding of 4th International Conference on Language Resources and Evaluation*, 2004.

[7] U. Quasthoff, M. Richter and C. Biemann, "Corpus Portal for Search in Monolingual Corpora," in *Proceeding of the 5th International Conference on Language Resources and Evaluation*, 2006.

[8] D. Broeder, R. van Veenendaal, D. Nathan and S. Stromqvist, "A Grid of Language Resource Repositories," in *Proceeding of 2nd IEEE International Conference on e-Science and Grid Computing*, 2006.

[9] T. Ishida, A. Nadamoto, Y. Murakami, R. Inaba, T. Shigenobu, S. Matsubara, H. Hattori, Y. Kubota, T. Nakaguchi and E. Tsunokawa, "A Non-Profit Operation Model for the Language Grid," in *Proceeding of the 1st International Conference on Global Interoperability for Language Resources*, 2005.

[10] D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Witterburg and C. Zinn, "A Data Category Registry- and Component-based Metadata Framework," in *Proceeding of 7th International Conference on Language Resources and Evaluation*, 2010.

[11] D. Van Uytvanck, C. Zinn, D. Broeder, P. Witterburg and M. Gardellini, "Virtual Language Observatory: the Portal to the Language Resources and Technology Universe," in *Proceeding of 7th International Conference on Language Resources and Evaluation*, 2010.

[12] B. Jorg, H. Uszkoreit and A. Burt, "LT World: Ontology and Reference Information Portal," in *Proceeding of 7th International Conference on Language Resources and Evaluation*, 2010.

[13] Y. Hayashi, T. Declerck, N. Calzolari, M. Monachini C. Soria and P. Buitelaar, "Language Service Ontology," The Language Grid, pp. 85-100, 2011.

[14] N. Calzolari, C. Soria, R. Del Gratta, S. Goggi, V. Quochi, I. Russo, K. Choukri, J. Mariani and S. Piperidis, "The LREC 2010 Resource Map," in *Proceeding of 7th International Conference on Language Resources and Evaluation*, 2010.

[15] H. Tohyama, S. Kozawa, K. Uchimoto, S. Matsubara and H. Isahara, "Construction of a Metabara Database of Efficient Development and Use of Language Resources," in *Proceeding of 6th International Conference on Language Resources and Evaluation*, 2008.

[16] S. Weibel, J. Kunze, C. Lagoze and M. Wolf, "Dublin Core Metadata for Resource Discovery," RFC 2413, The Internet Society, 1998.

[17] S. Kozawa, H. Tohyama, K. Uchimoto and S. Matsubara, "Automatic Acquisition of Usage Information for Language Resources," in *Proceeding of the 6th International Conference on Language Resources and Evaluation*, 2008.

[18] Miller, George A., "WordNet: A Lexical Database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.


**Shunsuke Kozawa** received the master's degree in Information Science from Nagoya University in 2009. Currently, he is a Ph.D. student at Nagoya University. His research interests include natural language processing and information retrieval. He is a member of the IEICE, the IPSJ and the ANLP.

**Hitomi Tohyama** received the master's degree in Language and Culture and the Ph.D. degree in Information Science from Nagoya University in 2004 and 2007, respectively. She was a Researcher from 2007 at Nagoya University. Her research interests include cognitive science, language usage and corpus analysis.

**Kiyotaka Uchimoto** received the B.E. and M.E. degrees in electrical engineering, and the Ph.D. degree in informatics, from Kyoto University, Kyoto, Japan, in 1994, 1996, and 2004, respectively. He is a Senior Research Scientist of the National Institute of Information and Communications Technology. His main research area is corpus-based natural language processing, and he specializes in Japanese sentence analysis and generation and information extraction. He is a member of the ANLP, the IPSJ, and the ACL.

**Shigeki Matsubara** received the M.E. degree and the Dr. of Engineering from Nagoya University in 1995 and 1998, respectively. He is an Associate Professor at Graduate School of Information Science, Nagoya University. His research interests include natural language processing, information retrieval and digital library. He is a member of the IEICE, the IPSJ, the ANLP, the IEEE and the ACM.