言語資源メタデータデータベースSHACHIの構築と利用

遠山 仁美† 小澤 俊介† 内元 清貴‡ 松原 茂樹† 井佐原 均‡

†名古屋大学 ‡情報通信研究機構

1 はじめに

音声・言語に関する研究開発での利用を目的に、コーパス、辞書、シソーラス等、様々な言語資源の開発が国内外の研究機関で進められ、その蓄積・利用環境の整備が進みつつある。欧米では、OLAC, ELRA, LDC など、言語資源の流通基盤の形成が組織的に進められているのに対して、日本国内では、そのような活動は必ずしも十分ではない。言語資源の開発には膨大なコストを要することを鑑みると、今後は、言語資源の共有及び再利用の仕組み作りが極めて重要となる [1]。

本稿[1] では、情報通信研究機構 (NICT) と名古屋大学が共同で開発を進めている、大規模言語資源メタデータデータベース SHACHI[2] について述べる。SHACHI では、世界、特に、アジア諸国の言語資源を対象に、詳細なメタ情報ならびに言語資源の関係を記述している。メタデータは、OLACmetadataSet を拡張したものであり、新たに 19 項目を追加している。データの品質を確保するために、情報収集及びデータ登録は専門家により実施している。SHACHI に収録されている言語資源はすでに 1700 件を超え、また、メタデータを活用した検索機能を装備するなど、言語資源の流通拠点としての役割の遂行が期待される。

2 SHACHIの設計

SHACHI を構築する目的は以下の通りである。

1. **言語資源メタデータの網羅的蓄積**：世界中の言語資源を対象に詳細なメタ情報を半自動により収集・作成する。
2. **言語資源間の関連付け**：詳細なメタ情報の収集によって言語資源を特徴付けることにより、言語資源間の関係が明示化され、言語資源オントロジー [4] としての提供も可能となる。
3. **言語資源開発状況に関する調査・分析**：作成したメタデータデータベースを網羅的に調査することにより、言語資源のタグ付け状況の実態や言語資源の種類の傾向などの把握が可能となる。
4. **言語資源の流通促進**：言語資源メタデータデータベースに検索機能を整備し、ユーザのニーズに合致した言語資源へのアクセスを容易にすることにより、言語資源の効率的利用を可能とする。

言語資源の蓄積・流通のための組織として、欧米では、Linguistic Data Consortium(LDC)、European Language Resources Association(ELRA)、Open Language Archives Community (OLAC) などのコンソーシアムがあり、日本国内には言語資源協会 (GSK) が設立され、いずれも重要な役割を果たしている。一方、言語資源のメタ情報を体系的に蓄積する試みとしては、DFKI[2] が運営する Language Technology World [3]、OLAC が提供する Web サイト[4]、などが挙げられる。

一般に、情報技術の進展ならびに社会への還元を促進する上で、、複数のコンソーシアムが相互に連携して活動することが重要である。SHACHI のメタデータは、OLAC のメタデータセットに準拠しており、それを拡張する形で、より詳細なメタ情報を収集している。これは同時に、ダブリン・コア (Dublin Core) のメタデータに準拠していることを意味しており、メタデータの蓄積・流通に適した設計となっている。

表 1: SHACHI のメタデータセット

	DC Qualifiers		Qualifiers for more precise description of the resources	
	LEVEL 1		LEVEL 2	
	Element	Qualifier	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 (4) corpus dictionary thesaurus glossary
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			

3 メタデータの収集

表 1 に SHACHI のメタデータセットを示す。本データベースの言語資源メタデータは、ダブリンコアの 15 の基本エレメント (表 1 左端欄を参照) に基づく OLAC

[1]SHACHI: A Large Scale Metadata Database of Language Resources –Construction and Utilization–
Hitomi Tohyama, Shunsuke Kozawa(Nagoya University), Kitotaka Uchimoto (NICT), Shigeki Matsubara(Nagoya University) Hitoshi Isahara (NICT)

[2]http://www.dfki.de/lt/publications_show.php?id=148

[3]http://www.lt-world.org/

[4]http://www.language-archives.org/

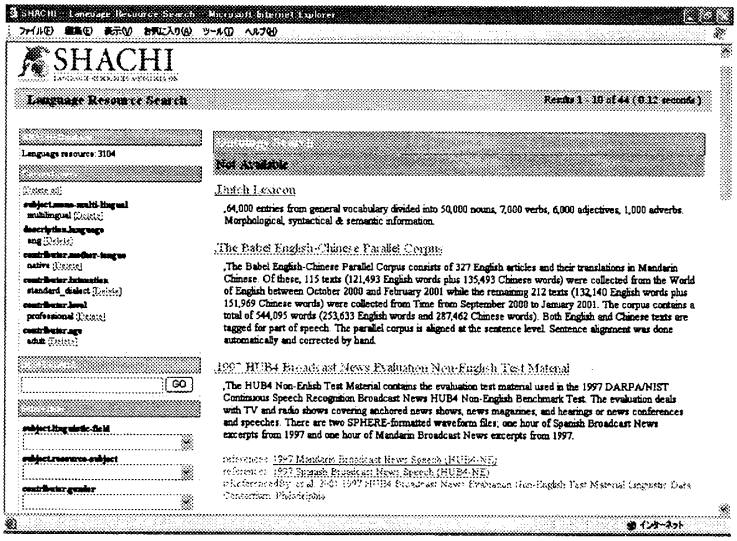

図 1: SHACHI の検索インタフェース

メタデータセットに準拠しており、さらに言語資源の特徴記述に必要となるメタデータ項目を追加している [3]。表 1 右端の欄は SHACHI 独自の拡張項目を示している。

SHACHI で収集する言語資源は、下記の全ての条件を満たすものと規定している。

1. デジタル化された言語資源である。
2. コーパス、辞書、シソーラス、語彙リストのいずれかである (数値のみのデータは、対象としない)。
3. 英語で記載された公式 Web ページを有し、かつ、データが公開されている。
4. 研究機関、研究者、企業によって作成された言語資源である。

言語資源の有機的結合に関する研究や、言語資源の流通促進、言語資源の戦略的開発を行うためには、世界中の言語資源のメタ情報が一ヶ所に網羅的に蓄積されていることが有用である。SHACHI では、国内の主要言語資源コンソーシアムを始め、欧米の言語資源コンソーシアムの持つメタ情報をカバーするとともに、SHACHI のメタデータセットに合わせた、より詳細なデータの人手による登録を実現している。

また、言語資源メタデータの収集においては、広く認知され、頻繁に利用される言語資源の情報を重点的に収録しておくことが重要である。SHACHI では、ELRA、OLAC といった主要コンソーシアムに登録されており、かつ、Web 検索で上位にランキングされる言語資源も随時調査し、登録している。

4 SHACHI の検索機能

SHACHI は、カタログリスト、個々の言語資源カタログ、SHACHI に収録された言語資源メタデータの統計情報、及び、言語資源検索ツールから構成されている。

検索ツールは、SHACHI のサイトを訪れたユーザが、目的に合った言語資源カタログに容易にアクセスできるように、キーワード検索、ファセット検索という 2 つのタイプの検索機能を備えており、収録されている言語資源を多様な角度から検索できる。図 1 に検索インタフェースの画面を示す。キーワード検索では、ユーザーがキーワードを自由に入力し、SHACHI のメタデータアーカイブの全単語を対象に検索できる。また、ファセット検索では、SHACHI が収集しているメタ項目から主要項目を抜粋し、それを選択項目として設定している。ユーザは、自分の希望する言語資源に近い項目を順に選択し、絞り込むことにより、該当する言語資源にたどり着ける。

5 まとめ

本稿では、言語資源データベース SHACHI について、設計、メタデータ拡張、構築、及び、検索機能について述べた。SHACHI は、広く認知されている言語資源を網羅的に収録している。ELRA、LDC といった言語資源コンソーシアムが提供する言語資源メタ情報をカバーしており、より詳細なメタデータが登録されている。現在、約 1700 件の言語資源メタ情報を収集しており、世界最大規模の言語資源メタデータアーカイブとなっている。

SHACHI の特徴の 1 つに、言語資源に付与されているタグセットやフォーマットのサンプルを収集していることが挙げられる。このため、言語資源の戦略的開発の策定（タグの標準化、言語資源間の有機的結合）に適している。また、メタ項目セットの小分類項目においては、極めて詳細なメタ情報を人手によって入力している。現在、それらの情報を用いて、各言語資源の近さを計測し、世界の言語資源の体系的な蓄積・整理を試みている。

参考文献

[1] T. Ishida, et al.: A Non-Profit Operation Model for the Language Grid, *Proceedings of ICGL2008*, pp.114–121 (2008).

[2] `http://www.shachi.org/`

[3] H. Tohyama, et al.: SHACHI: A Large Scale Metadata Database of Language Resources, *Proceedings of ICGL2008*, pp.205-212 (2008).

[4] Y. Hayashi, et al.: Ontologies for a Global Language Infrastructure, *Proceedings of ICGL2008*, pp.105-112 (2008).