言語資源メタデータデータベースSHACHIの構築

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
{hitomi, kozawa, matubara}@el.itc.nagoya-u.ac.jp

‡情報通信研究機構
{uchimoto, isahara}@nict.go.jp

1. はじめに

近年、音声・言語の主要メディアに関する研究開発を目的に、電子化されたコーパス、辞書、シソーラスといった言語資源（音声・動画を含む）の構築が盛んに行なわれ、その重要性は広く認識されている。しかし、これまでに公開されている言語資源は、開発機関において、個々の目的に応じて構築されており、タグセットやフォーマットにおいても多種多様で、それぞれの言語資源がほぼ独立に存在している。しかし、言語資源の構築段階において、他の言語資源の仕様に準拠したり、同じタグセットを参照するなど、実際には、個々の言語資源間に何らかの関連性があることも少なくない。

欧米では、Linguistic Data Consortium (LDC)、European Language Resources Association (ELRA) といった言語資源コンソーシアムが、主に欧米語の言語資源の収集、配布を行っている。また、Open Language Archives Community (OLAC)では、言語資源のメタデータの統一や、言語資源カタログの整備を担っている。しかし、カタログの情報は詳細ではなく、それらの言語資源が研究開発過程のどのようなフェーズで用いられたかといった情報や、他の資源との関係性などは示されていない。そのため、流通という面で十分機能しているとは言えない。国内においても、国立情報学研究所(NII)や、言語資源協会（GSK）によって、音声・テキストコーパスを蓄積する活動が行われている。しかしながら、体系的に蓄積するには至っておらず、このような状況において、多様な目的を持ったユーザが、目的に合致する言語資源にたどり着くことは容易ではない。

そこで、情報通信研究機構(NICT) と名古屋大学は共同で、大規模言語資源メタデータデータベースSHACHI の構築を 2007 年から開始した。SHACHI は、日本・アジア諸国の言語資源をはじめ、世界中の言語資源の詳細なメタデータを収集し、得られた知見を基に体系的な蓄積を試みている。SHACHI のメタデータセットは、OLAC のメタデータセットを拡張したものであり、新たに 19 項目を追加している。SHACHI

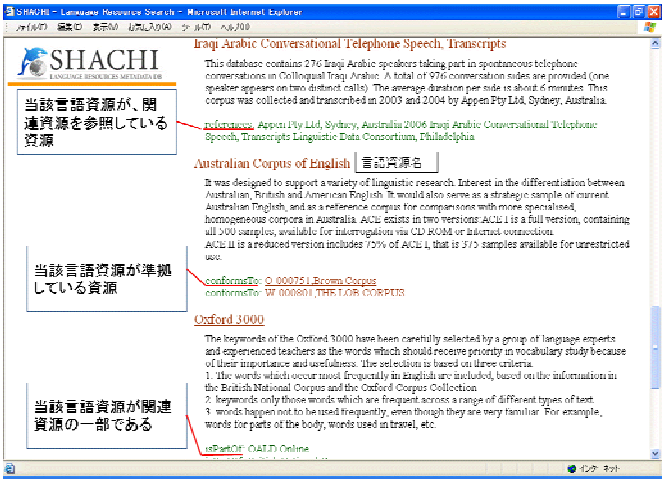

図 1. 言語資源間の関係性の明示化
(SHACHI 検索結果の表示)

に収録されている言語資源はすでに 1800 件を超え、また、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が期待される。

本稿では、SHACHI の目的、設計、メタデータの収集・拡張、及び、カタログ検索機能について述べる。

2. SHACHI の目的

言語資源メタデータデータベースSHACHIを構築する目的は以下の 5 点にまとめられる。

(1) 言語資源メタデータの蓄積: 既存の言語資源を有機的に結合したり[1]、戦略的に言語資源を開発するためには、世界中の言語資源に関する情報が一個所にまとめられていることが重要である。そこで、大量の言語資源の詳細なメタデータを半自動で収集し、各言語資源の詳細なカタログを作成する。

(2) 言語資源メタデータの体系化: (1)によって収集された詳細なメタデータをもとに、言語資源のタイプを分類することにより、言語資源オントロジーの構築を試みる[2]。図 2 は、開発中の言語資源オントロジーの例である。現段階では、人手によって構築しているが、オントロジーを自動生成する手法の考案を試みている。

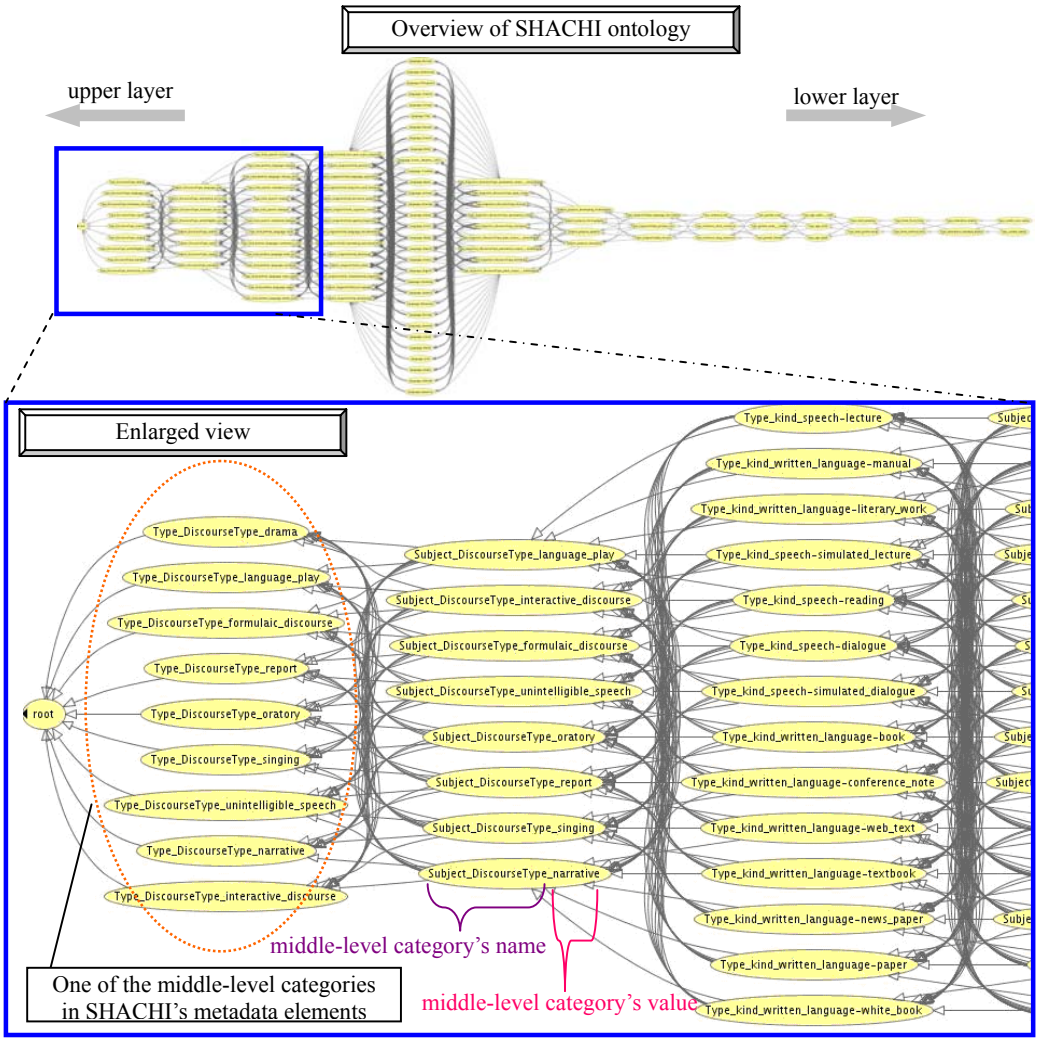

図 2. 言語資源オントロジーの構築 （人手による構築の例）

(3)言語資源間の関連付け： SHACHI は LDC、ELRA のカタログなどに比べ、詳細なメタデータを収集している点に特徴がある。これらの詳細なメタデータは、個々の言語資源の特徴の記述を可能にし、さらには、言語資源間の関係性を明示化することが期待できる。図 1 は、SHACHI の検索画面の一部である。検索結果として表示された言語資源が準拠した資源（参照元）や、フォーマットが共通である資源を提示している。これらの言語資源間の関係性を統計的に調査することにより、世界標準レベルのタグセットや、データフォーマット、需要のある言語資源のタイプなどを策定することができる。

(4)言語資源の統計調査： SHACHI のサイトでは、収録されている言語資源メタデータに関する統計情報が閲覧できる。これらメタデータを統計的に分析することで、どのような資源が世界のどこに存在するのかを把握したり、近年、公開されている言語資源の傾向を捉えることができる。

(5)言語資源の流通促進： 本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、言語資源の有効利用や、効率的開発を支援する。

3. SHACHI の設計

言語資源のメタデータを体系的に蓄積したり、アクセス性を高める試みとしては、上述したOLAC以外にも、ISLE Meta Data Initiative（IMDI）、DFKIが運営するLanguage Technology World 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、より詳細なメタデータを収集している。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。また、現在、言語属性を示すメタ項目は、国際標準化機構（ISO）のISO639-2[3]に準拠している。また、日付と時刻の表記に関してはISO8601に準拠している[4]。

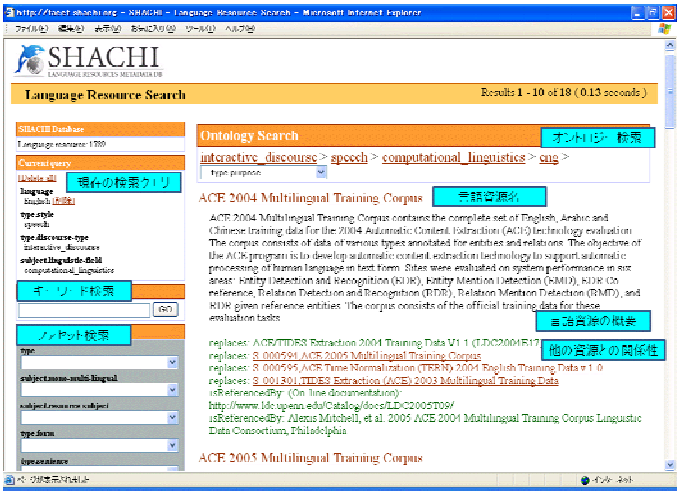

図 3.　SHACHI の検索画面　（http://www.facet.shachi.org/）

4. 言語資源メタデータの収集

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。

a. デジタル化された言語資源である。
b. コーパス、辞書、シソーラス、語彙リストのいずれかである。
c. 英語で記載された公式 Web ページ、もしくは、言語コンソーシアム管理下のカタログを有し、かつ、データが公開されている。
d. 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。

これらの条件を踏まえた上で、認知度が高いもの、言語処理技術の発展の上で、主要であると考えられるもの、大規模であるものを優先的に登録している。また、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、世界の言語資源のメタデータを網羅的に一箇所に集めておくことが有効である。そこで、SHACHI では、国内の主要言語資源コンソーシアムを始め、欧米、中国の言語資源コンソーシアムの持つ言語資源メタデータをカバーしている[5]。

5.メタデータ拡張

ユーザの曖昧な検索に対し、必要十分な情報を提供するためには、言語資源の属性を示す詳細なメタデータに加え、言語資源の属性の近さや上位・下位関係を体系的に表示したオントロジーを構築することが必要であると考えられる。本データベースは、ダブリン・コアの15の基本エレメントに基づくOLAC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の記述に必要と判断した、新たなメタデータ19項目を追加している。表2の編みかけ部分はSHACHIが採用しているメタ項目であり、左から、ダブリン・コア、OLAC、右端はSHACHI 独自の拡張項目を示している。また、個々の言語資源が、研究者らによって、どのような局面で、どう活用されたかという用途情報は、ユーザにとって貴重である。そこで、本研究では、構文解析技術を用い、言語資源の用途に関する情報を、学術論文から自動抽出する手法を考案し、提供を試みている[6]。

6. SHACHI のカタログ検索

本メタデータデータベースのユーザが、目的に合った言語資源カタログに到達できるよう、キーワード検索、及び、ファセット検索の2つの機能を整備している。また、現段階では、人手によって構築した言語資源オントロジーを用い、オントロジー検索機能を試験的に搭載している。図 3 に、検索ツールの画面を示す。ファセット検索機能は、SHACHI メタデータセットから 15 種類の主要メタ項目が選択項目として設置されており、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、絞り込んでいくことで、該当する言語資源にたどり着くことができる。一方、オントロジー検索では、言語資源オントロジーをたどる方法により、ユーザのニーズに合致する言語資源に到達するシステムであり、漠然とした目的を持ったユーザや、初心者、ならびに、一般のユーザにとって有効であると考えられる。

表2. SHACHIのメタデータセット

		Qualifier		
		DC Qualifiers	Qualifiers used for more precise description of the resources	
		LEVEL 1	LEVEL 2	
	DCMES Elem	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		

7. まとめ

情報通信機構(NICT)と名古屋大学では、共同で、言語資源の有効利用、戦略的開発、及び、有機的結合に関する研究を目的に、言語資源のメタデータを大規模に収集している。本稿では、言語資源メタデータデータベース“SHACHI”の設計、メタデータの収集・拡張、及び、検索機能の実現について述べた。SHACHI は Web 検索によって、認知度の高い言語資源を網羅的に収録しており、また、世界の主要言語資源コンソーシアムが提供している言語資源メタデータをカバーするとともに、さらに、より詳細なメタデータの登録を人手により行っている。現在、約 1800 件の言語資源メタデータの登録を完了し、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、極めて詳細なメタデータを収集している点があげられる。現在、それらの情報を用いて、言語資源のタイプや各言語資源間の近さ(属性の近さ)を計測し、世界中の言語資源メタデータの体系的な蓄積(言語資源オントロジーの構築)を試みている。

謝辞

本データベース構築にご尽力頂きました、北陸先端科学技術大学院大学の白井清昭先生、NICT 登録スタッフの前川絵美さん、翻訳家の和氣祥子さん、名古屋大学文学研究科の大西美穂さん、同情報科学研究科松原研究室の小野貴博さん、杉木健二さん、(株)アンカーの登録スタッフの皆様に感謝申し上げます。

参考文献

[1] 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, ICGL, pp.114-121 (2008).

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

[3] H. Tohyama, S. Kozawa, K. Uchimoto, S. Matsubara, H. Isahara, SHACHI: A Large Scale Metadata Database of Language Resources, ICGL, pp.205-212 (2008).

[4] ISO639-2: Codes for the representation of names of languages -- Part 2: Alpha-3 code.

[5] ISO8601: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=4087

[6] 小澤俊介・遠山仁美・内元清貴・松原茂樹, 言語資源の効率的利用のための用途情報抽出,言語処理学会第14回年次大会発表論文集, (2008).