異文化間コミュニケーション支援のための言語資源の連携に関する研究 (継続)

研究代表者　林　良彦　大阪大学大学院言語文化研究科教授
共同研究者　楢和千春　京都工芸繊維大学非常勤講師

1 はじめに

社会のグローバル化により，母国語が異なる人々の間でコミュニケーション(異文化間コミュニケーション) の機会が増大している．また，情報技術の進展，ネットワークの普及に伴い，情報通信技術(ICT)を活用した異文化間コミュニケーションが現実的なものになっている．異文化間コミュニケーションの支援には，様々なレベルや形態が考えられるが，とくに言語コミュニケーションに対する支援の必要性は高く，機械翻訳に代表される自然言語処理技術や辞書やコーパスといった言語資源を効率的に組み合わせて利用することを可能とする言語基盤 (language infrastructure) への期待が高まっている (林, 2007)．本研究では，異文化間コミュニケーション支援を目的とした言語資源の連携に関する研究を行う．

今期は特に，セマンティック Web の基盤の上に構築される言語基盤における辞書言語資源を組み合わせて利用するというシナリオにおいて必要となる以下の 2 点について重点的に研究を進めた．

- 辞書言語資源アクセス機能のオントロジー化 (3 節)
- 複合的な語彙資源アクセスのための辞書言語資源のモデル化 (4 節)

2 言語サービス基盤と言語サービスオントロジー (Hayashi et al., 2008a)

さまざまな言語資源や言語処理ツール・システムが公開され利用可能となっていること，また，いわゆる Web サービスに関する技術が普及してきたことにより，Web 上の言語基盤(language infrastructure)を構築しようとする動きが活発化している．報告者がかかわっている独立行政法人・情報通信研究機構による言語グリッドプロジェクト (http://langrid.nict.go.jp) もその活発な一例であり，本プロジェクトでは異文化コラボレーションの支援を目的としている．

言語サービスの構成要素となる言語資源や言語処理ツールは，独自の目的のために独立に構築されたものが多く，その再利用性や相互運用性に関しては共通する技術的な課題をかかえている．たとえば言語資源データについては，データフォーマットや言語的注釈のタグ体系が固有のものであることが多い．また言語処理ツールについては，入出力データやアクセスメソッドがさまざまである．このような言語資源や言語処理ツールの独自性(idiosyncrasy) を隠蔽し，互いを整合させるためのひとつの考え方として，言語基盤上の構成要素の単位を原始的な Web サービス(atomic Web service)と考え，これらに対して標準的なアクセス手段(API)を規定することが考えられる．ここで，標準的なアクセス手段である Web API の体系は，共通的に理解される意味的な基盤に基づいていることが望まれる．報告者らは，海外の関連研究機関と連携し，言語サービスの構成要素に対する意味的な基盤を与えるための言語サービスオントロジーの体系化を進めてきた(Hayashi, 2008a)．

図 1 に本研究代表者らが提案する言語サービスオントロジー(Hayashi, et al. 2008a)の最上位階層を示す．言語サービス(**LanguageService**)は言語処理資源(**LanguageProcessingResource**)により提供される(**providedBy**)．言語処理機能は言語データ資源(**LanguageDataResource**)を利用し，言語表現(**LinguisticExpression**)，すなわち言語データを処理する．また，ひとつの言語表現は，多重の言語的注釈(**LinguisticAnnotation**)により注釈付けられる．これにより，さまざまなレベルの言語解析結果や複数の言語解析器の結果を対象の言語データと関係付けられる．図 1 における各ボックスはそれぞれが独立したクラスであり，さらにサブオントロジーとして詳細化される．本報告では，図 1 における言語データ資源クラスの一部をなす辞書言語資源について論じる．グローバルかつオープンな言語基盤においては，言語サービスオントロジーは広く関係者に共有されている必要があり,最終的には何らかの標準化が必要となる．言語サービスオントロジーの標準化へ向けては，部分的にでも関連する国際標準がすでに存在する場合，そ

れらを適切に利用する，あるいは，取り込んでいくことが必要となる．その際は，国際標準の規格・仕様をオントロジー化（ontologized）することが必要となる．

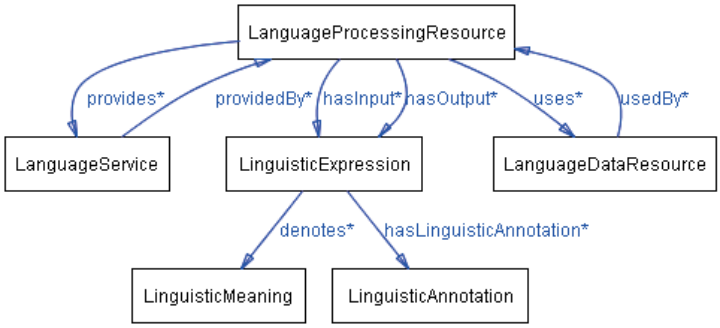

図 1: 言語サービスオントロジーの最上位階層

3 辞書言語資源のアクセス機能のオントロジー化 (Hayashi et al., 2008b)

昨年度の検討において，辞書言語資源のオントロジー化の提案を行った．その基本的な考え方は，辞書言語資源のメタモデルとして国際標準化された[1]LMF（Lexical Markup Framework）(Francopoulo et al. 2006) をオントロジー化することに基づいている．ここで，オントロジー化とは， UML で与えられている LMF のデータモデルをオントロジー記述言語である OWL へと変換することを意味する．

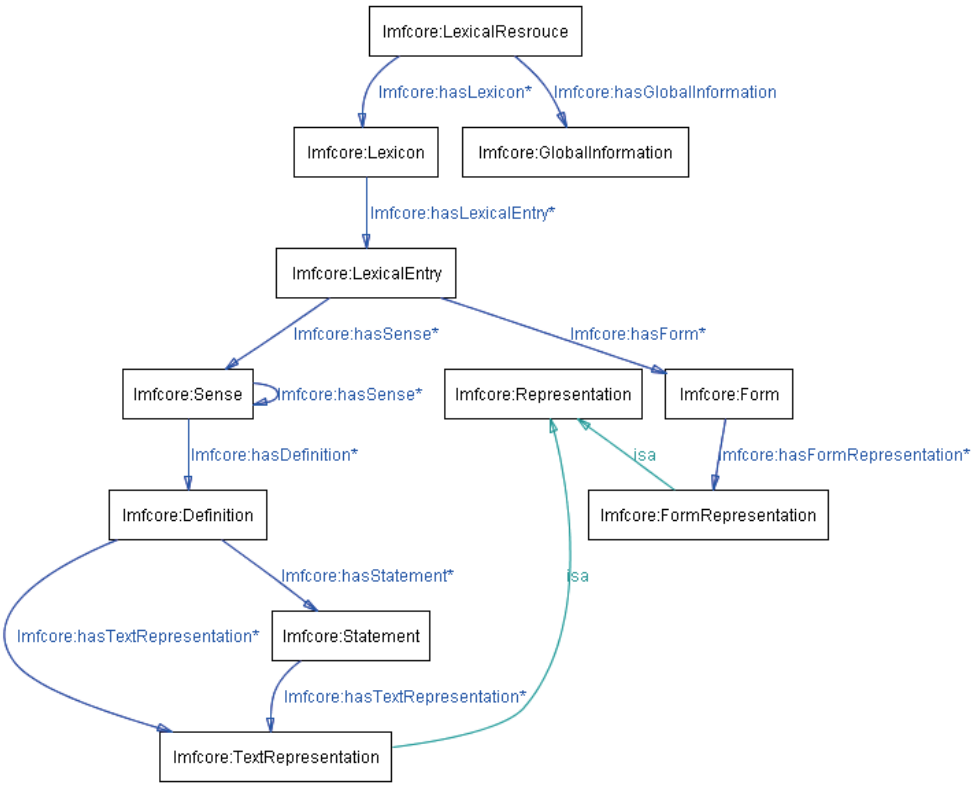

図 2: LMF Core Model のオントロジー的規定

図 2 に LMF の中核部分である Core Model に対応する OWL 記述の主要な部分を図式化したものを示す．辞書(**Lexicon**)は，複数の辞書エントリ(**LexicalEntry**)を持つこと，各辞書エントリは語形(Form)に関す

[1] ISO-24613:2008

る情報と意味に関する情報(Sense)を持つこと，これらはテキスト(**TextRepresentation**)によって表わされることなどが規定されている．

ところで，辞書言語資源には，人間用の辞書，計算機処理(自然言語処理)用の辞書といった利用主体に基づく分類があり，さらに，例えば人間用の辞書であれば，モノリンガル辞書(国語辞書など)や対訳辞書などといったように細分類される．我々は，辞書のタイプがそのエントリ項目の構造などにより規定されることに注目し，図 3 に示すような辞書言語資源のタクソノミーを提案する．ここで，各タイプの辞書のエントリ構造は，メタモデルである LMF の枠組みで記述された辞書エントリの規定と結び付けることにする．このようにすることで，辞書言語資源のモデルに関する国際標準と適合した言語サービスオントロジーのサブオントロジーの開発を行うことができる．

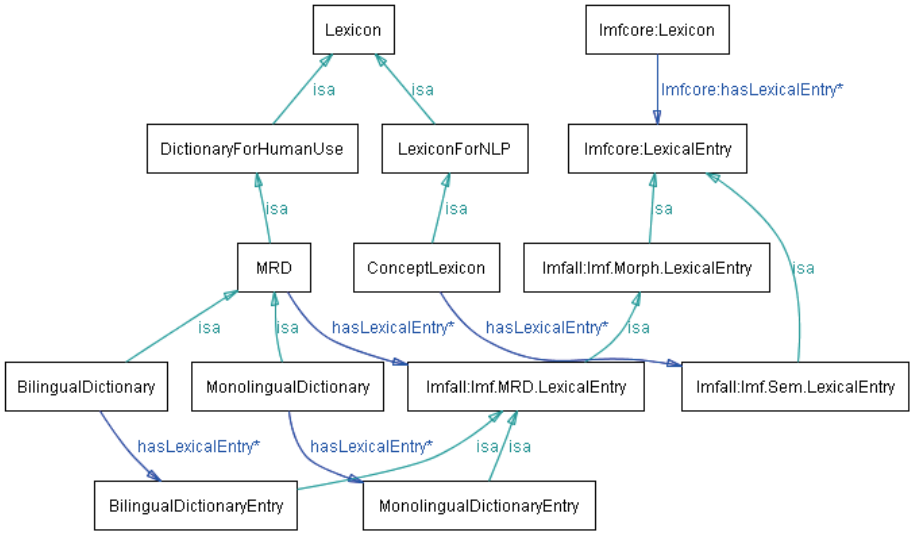

図 3: 辞書言語資源のタクソノミーと LMF オントロジーとの対応付け

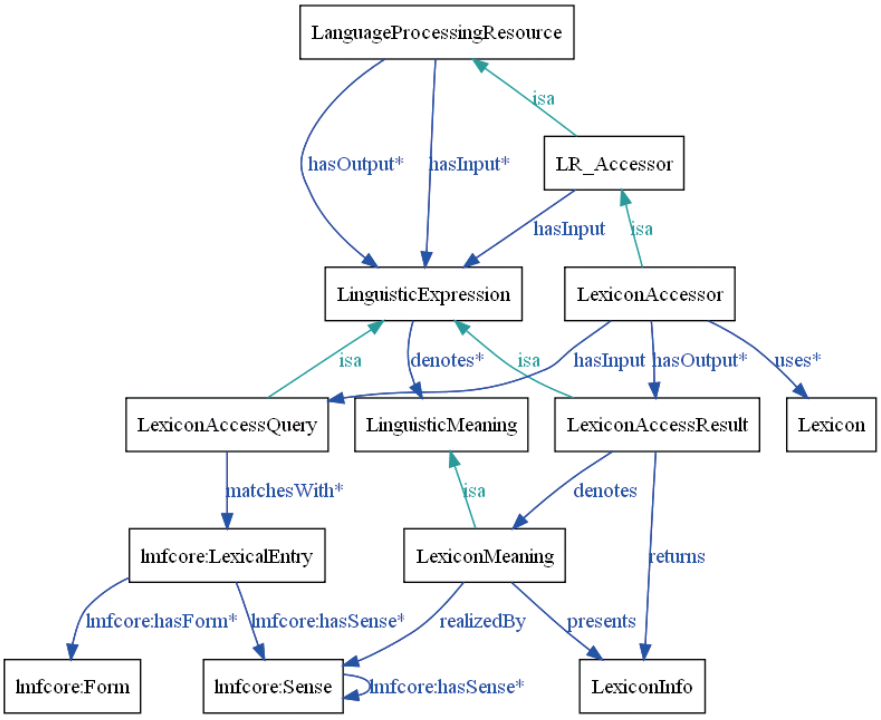

図 4: 辞書言語資源アクセス機能のオントロジー的規定

図 4 に言語サービスオントロジーにおける辞書言語資源アクセス機能(**LexiconAccessor**)のオントロジ

一的規定を示す．すなわち，辞書言語資源アクセス機能は，言語処理資源(**LanguageProcessingResource**)の一種である言語資源アクセス機能(**LR_Accessor**)のサブクラスである．また，アクセスする見出し語の言語表現(**LinguisticExpression**)を入力とし，辞書アクセス結果（**LexiconAccessResult**）クラスに属するデータを出力する．このクラスは，言語表現クラスの下位クラスであり，辞書的意味(**LexiconMeaning**)クラスで表現される意味を指示する(**denote**)．また，辞書的意味は，LMF により規定される LMF 意義クラス(**lmfcore:Sense**)により実現される(**realizedBy**)．ここで，LMF 意義クラスは，辞書における意味情報をモデル化するクラスである．以上より，辞書言語資源のアクセス機能は，見出し語を入力(検索キー)とし，辞書に記述されている意味情報を出力するものして規定される．

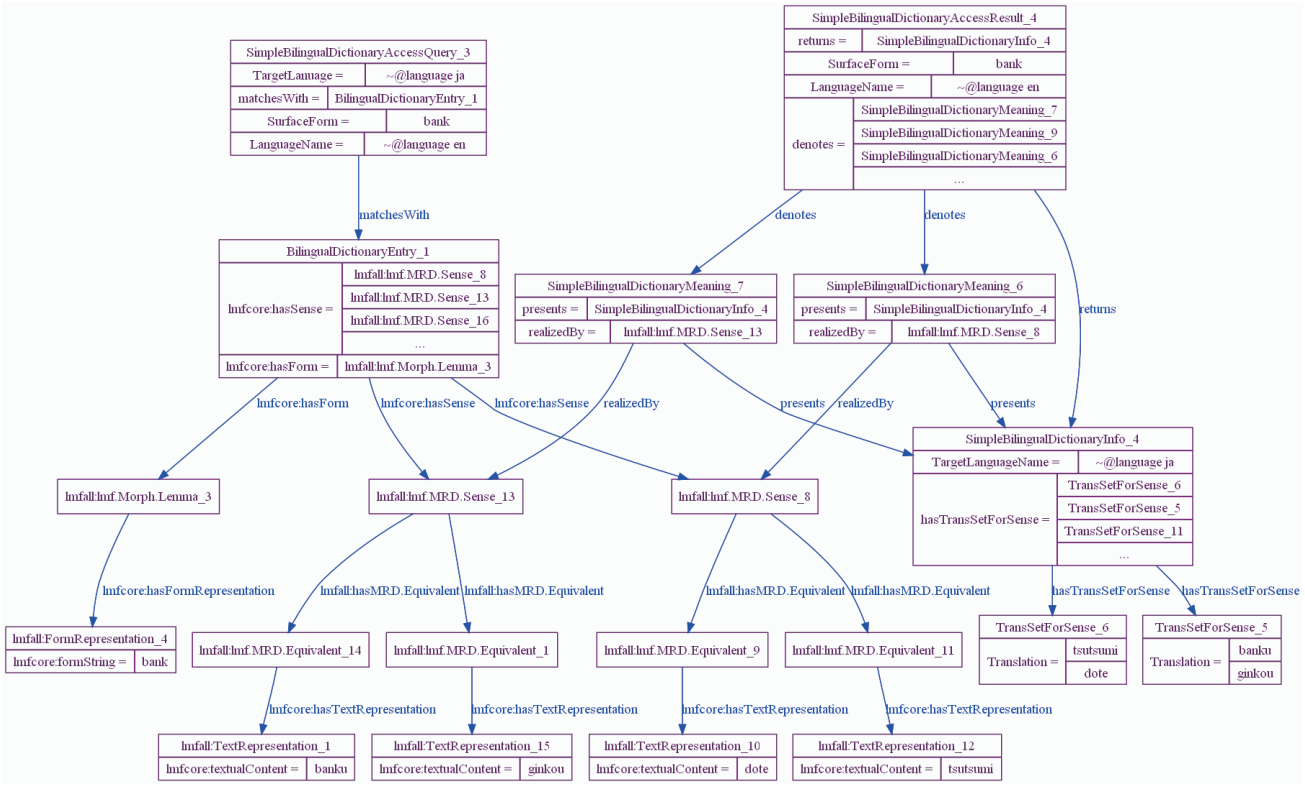

図 5: 具体的な辞書エントリの記述例

図 5 に英日対訳辞書における見出し語'bank'に関する具体例を示す．このような規定は依然として概念的なものであり，実際の Web サービスにおける API との対応付けが行えることが望ましい．より具体的には，Web サービスの API を規定する標準である WSDL（Web Service Description Language）記述と概念的なオントロジー体系との関連付けを行うことが望ましい．このような目的に沿う規格として SAWSDL（Semantic Annotations for WSDL）（SAWSDL，2007)と呼ばれるものがある．SAWSDL は WSDL のように標準化はなされていないが，W3C の勧告となっているもので，一定の支持コミュニティが存在する．SAWSDL においては，**sawsdl:modelReference** という要素を用いて，WSDL における入出力データの持つセマンティックスを背景にあるドメインオントロジーの概念クラスと対応付ける．図 6 に上記の単純な対訳辞書アクセス機能に対応する SAWDSL 文書の一部を示す．

```
.....
  <wsdl:portType name="SimpleBilingualDictionary">
    <wsdl:operation name="serachSimpleBilingualDictionary"
        sawsdl:modelReference="http://langrid.nict.go.jp/lso/lso#SimpleBilingualDictionaryAccessor">
      <wsdl:input message="sbd:serachSimpleBilingualDictionaryRequest"/>
      <wsdl:output message="sbd:serachSimpleBilingualDictionaryResponse"/>
    </wsdl:operation>
.....
    <xsd:element name="serachSimpleBilingualDictionaryRequest"
       type="sbd:SearchQuery">
    </xsd:element>
    <xsd:element name="serachSimpleBilingualDictionaryResponse"
       type="sbd:SearchResults">
    </xsd:element>
.....
    <xsd:complexType name="SearchQuery"
       sawsdl:modelReference="http://langrid.nict.go.jp/lso/lso#SimpleBilingualLexiconAccessQuery">
       <xsd:sequence>
              <xsd:element name="SurfaceForm" type="xsd:string"></xsd:element>
              <xsd:element name="LanguageName" type="xsd:language"></xsd:element>
              <xsd:element name="TargetLanguage" type="xsd:language"></xsd:element>
       </xsd:sequence>
    </xsd:complexType>
.....
    <xsd:complexType name="SearchResults">
       sawsdl:modelReference="http://langrid.nict.go.jp/lso/lso#SimpleBilingualDictionaryInfo">
       <xsd:sequence>
              <xsd:element name="Language" type="xsd:language"></xsd:element>
              <xsd:element name="TransSetForSense" type="sbd:TransSetForSenseType"
                     maxOccurs="unbounded" minOccurs="1"></xsd:element>
       </xsd:sequence>
    </xsd:complexType>
    <xsd:complexType name="TransSetForSenseType">
       <xsd:sequence>
              <xsd:element name="Translation" type="xsd:string" maxOccurs="unbounded" minOccurs="1"></xsd:element>
       </xsd:sequence>
    </xsd:complexType>
.....
```

図 6: 単純な対訳辞書アクセス機能の SAWSDL 文書の例

4 複合的な語彙資源アクセスのための辞書言語資源のモデル化 (林, 2009)

複合的な辞書を Web 上で仮想的に提供するための技術基盤を検討するために,EDR 電子化辞書 (EDR, 2003) の概念体系辞書を L1(日本語)辞書, Princeton WordNet (Fellbaum, 1998)を L2(英語)辞書, EDR 電子化辞書の対訳辞書を L1-L2 対訳辞書とする L1-L2-L2 辞書をケーススタディとし, その実現シナリオを検討する.

WordNet の情報構造は語彙化概念に基づいている. WordNet における基本的な情報単位は synset と呼ばれる同義語集合である. 単語はいくつかの語義を持ち, 各語義はある概念を指示する. ”car”, “automobile”のように異なる単語が共通の概念を指示する場合, これらの単語が synset を構成する. synset には自然言語テキストによる説明(gloss)が与えられる.

EDR 電子化辞書における特徴は, 日本語概念, 英語概念などの各辞書のすべてのエントリが日本語・英語にまたがる概念識別子によって関係付けられていることである. 概念識別子は, 日本語, 英語にまたがる(あるいは, 言語に依存しない)概念ノードを表し, 日本語, 英語による見出し語と概念説明テキストが付与される. たとえば, 0f74e9 という概念識別子はおおむね「自動車」の概念を表し, ”自動車”や”car”といった日本語, 英語の単語に関する各辞書のエントリはこの概念識別子と関係付けられている. このため, 概念識別子をキーとしてこれと関係付けられた単語の集合を求めることにより, 疑似的に日本語, 英語にまたがる synset を構成することができる. EDR 電子化辞書は, 形式的には Princeton WordNet と同様の情報構造を持ち, synset と gloss が日英両言語により与えられるものとして扱うことができる (Hayashi and Ishida, 2006).

複合的な語彙資源は, あらかじめ off-line の batch 的な処理により実現することもちろん可能である. 実際, 多くの語彙的オントロジーの対応付けに関する従来研究は, このような前提に基づいている. しかしな

がら，言語データ資源の Web サービス化は動的で仮想的な言語資源の実現の可能性をひらく．そこで，ユーザからの要求によって(on-demand)，動的に(on-the-fly)，複数の語彙資源を組み合わせアクセスすることにより，仮想的な複合辞書へのアクセス機能を提供するアクセスサービスの実現を考える．このようなアクセスサービスを実現するためには，異なる辞書のエントリ間の対応付けを行う必要がある．その方法は本研究の範囲ではないが，何らかの方法によりこれが可能であると仮定すると，異なる辞書のエントリ間の対応付けをユーザからの要求に即して機会主義的に(opportunistic) 行う一方，対応付けの結果を Web サービス基盤上に新たな言語資源として蓄積することが考えられる．EDR 電子化辞書や Princeton WordNet は，独立した一次的な言語資源である．これに対し，異なる辞書エントリ間の対応付けのデータは，二次的な言語資源と考えることができる．ところで，対応付けのような二次的な情報構造は，一次的言語資源の外部にあるべきである．すなわち，一次的言語資源には，二次的言語資源への参照は含まれるべきではない．これにより，一次的言語資源は独立に保ったまま，その上に二次的な言語資源を重畳させることが可能となる．

以下では，このような考え方に基づく実現を可能とする辞書言語資源のモデル化を示す．この検討は，前節でも用いた LMF の枠組みに基づき，これを拡張する．LMF では，その多言語拡張パッケージにおいて，異言語の辞書エントリの対応付けのために，**Sense Axis** と **Transfer Axis** という二つのクラスを導入している．これらはそれぞれ，機械翻訳方式におけるピボット方式とトランスファ方式に対応しており，**Transfer Axis** は，とくに二言語間の統語的な対応関係の表現に適している．一方，**Sense Axis** は意味的な対応関係を表すもので，とくに 3 つ以上の多言語間の等価的対応を表現するのに適している．

異なる辞書エントリ間の対応付けという二次的な言語資源を Web サービス化したものをここでは仮に Sense Axis サーバーと呼ぶ．Sense Axis サーバーは，関連付けられた特定の辞書エントリの組に関する情報も保持する必要がある．そこで，

- 対応関係：等価関係以外の語彙意味論的関係による対応付けがありえるため，その関係ラベルを保持する．
- 対応関係付与に関するメタ情報：いつ，どのようなプロセスにより，どの程度の信頼度をもって当該の関係が付与されたかを記録する．

といった情報を保持する **Sense Pair Relation** を導入し，Sense Axis サーバーにおいては **Sense Pair Relation** のインスタンスを対応する **Sense Axis** インスタンスと関係づけて集約することを提案する．**Sense Pair Relation** のインスタンスには，対応付けを保持している **Sense/Synset** インスタンスの ID と関連する **Sense Axis** インスタンスの ID を保持するとともに，対応付けにおけるメタ情報を保持する．図 7 にこれらのインスタンスの関係の概要を示す．

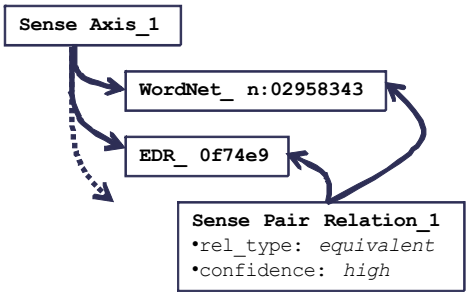

図 7：各クラスのインスタンス間の関係

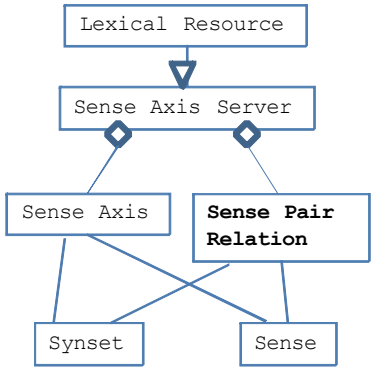

図 8：Sense Axis クラスの拡張

図 8 に LMF の **Sense Axis** クラスの拡張の概要を示す．ここでは，**Sense Axis Server** を **Lexical Resource** のサブクラスとして規定している．

Sense Axis サーバーの Web サービスの形態としていわゆる REST (Pautasso, 2008) を想定する．すなわち，サービスの機能と対応した一定の URI 形式によりサービスにアクセスし，結果データを XML 形式で受け取る．本研究のベースは LMF にあるので，結果データの形式は LMF の仕様に annex として提示されている XML 形式に準拠する．そこで焦点は，Sense Axis サーバーが有するべき機能と URI の設計となる．なお，WordNet や EDR などの語彙資源が Web サービス化されていることを前提とする．すでに，(Van Assem et al., 2006) は，REST 形式での WordNet サービスを提案している．

Sense Axis サーバーが有すべき基本機能は次のとおりであり，これらの機能との対応が明確な URI を割り当てる．

a. エントリ間の対応付けの検索と実行: 単語とターゲットとする語彙資源を与え，すでに辞書エントリ間の対応が得られていれば，その **Sense Axis** インスタンスの ID を返却する．対応付けが得られていなければ，動的な対応付け処理を行い，その結果を返す．
b. エントリ間の対応付け情報のアクセス: **Sense Axis** インスタンスの ID から関係付けられた **Sense/Synset** インスタンスの ID のリストを返却する．ここで，これらのインスタンスは各語彙資源における具体的なエントリであり，特定の URI 形式によりアクセスできるものとする．また，その辞書エントリの内容は，LMF/XML で得られるものとする．さらに，**Sense/Synset** インスタンスの組に関する **Sense Pair Relation** のインスタンスの ID のリストを返却する．
c. エントリ間の対応付け情報の登録: a の API による動的な対応付けの結果を適切なメタ情報とともに **Sense Pair Relation** のインスタンスに登録する．

5 おわりに

本研究では，セマンティック Web の基盤の上に構築される言語基盤において，複数の辞書言語資源を組み合わせて利用するというシナリオにおいて必要となる以下の 2 点について重点的に研究を進めた．

- 辞書言語資源アクセス機能のオントロジー化: ISO 国際標準 LMF に基づき，辞書言語資源をそのエントリが持つ特性によりサブクラス化し，タクソノミーを構成することを提案した．ちなみに，オリジナルの LMF では辞書クラスのサブクラス化を許していないが，これは個々の辞書のモデル化を目的としており，言語基盤というより広い視野の中でその構成要素である辞書言語資源のタクソノミーを構築するということは範囲外になっていることによる．また本検討では，具体的な Web サービスの API 規定と言語サービスオントロジーの概念的対応を SAWSDL の規格により対応付けることを提案した．この点に関しては，言語グリッドのような具体的な言語基盤での必要性や実現性の観点から，より詳しく検討していく必要がある．
- 複合的な語彙資源アクセスのための辞書言語資源のモデル化: 特に異言語・異体系の意味・概念辞書を動的に対応付け，この対応関係を Web サービス基盤上で二次的な言語資源として蓄積する枠組みを検討し，辞書モデリングに関する ISO 国際標準 LMF の適用と拡張について議論した．今後は，Web API における URI 形式の詳細と結果の XML 形式を定め，実際に Web サービスの実現を行う．また，異なる言語の異なる辞書エントリを on-the-fly/on-demand で対応付けるための効率の良い手法を追求していく必要がある．

なお，2 年間の助成を受けた本研究テーマ「異文化間コミュニケーション支援のための言語資源の連携に関する研究」は本年度で終了するが，その成果は 2009 年度から新たに開始される以下の二つの研究プロジェクトに引き継がれる．これまでの研究助成に対して，深く感謝する次第である．

- 総務省 SCOPE 受託研究: サービスコンピューティングに基づく多言語サービス基盤の実現（研究代表者: 石田 亨．研究分担者: 林 良彦，楠和千春ほか）
- 科学研究費補助金 基盤研究(C): 語彙資源を用いた概念の語彙化の分析・記述に関する研究（研究代表者: 林 良彦）

【参考文献】

1. 林 良彦. (2007). セマンティック Web と言語技術・言語資源. 情報処理, Vol.48, No.8, pp.857-863.
2. 林 良彦. (2009). 複合的な語彙資源アクセスサービスの実現基盤. 言語処理学会第 15 回年次大会発表論文集, pp.72-75.
3. Van Assem, M., Gangemi, A., and Schreiber, G. (2006). Conversion of WordNet to a standard RDF/OWL representation. *Proc. of LREC2006*.
4. EDR. (2003) EDR Electronic Dictionary Technical Guide. http://www2.nict.go.jp/kk/e416/EDR
5. Fellbaum, C. (Eds.). (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.
6. Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. (2006). LMF for Multilingual, Specialized Lexicons. In: *Proc. of LREC2006*, pp.233-236.
7. Hayashi, Y., Ishida, T. (2006). A Dictionary Model for Unifying Machine Readable Dictionaries and Computational Concept Lexicons. *Proc. of LREC2006*, pp.1-6.
8. Hayashi, Y., Declerck, T., Buitelaar, P., and Monachini, M. (2008a). Ontologies for a Global Language Infrastructure. In: *Proc. of ICGL2008*, pp.105-112.
9. Hayashi, Y., Narawa, C., Monachini, M., Soria, C., and Calzolari, N. (2008b). Ontologizing Lexicon Access Functions based on LMF-based Lexicon Taxonomy. In: *Proc. of LREC2008*, pp.231-237.
10. Pautasso, C. et al. (2008). RESTful Web Services vs. Big Web Services: Making the Right Architectural Decision. *Proc. of WWW2008*.
11. SAWSDL 2.0. (2007). Semantically Annotations for WSDL and XML Schema. http://www.w3.org/TR/sawsdl.

〈発 表 資 料〉

題　　名	掲載誌・学会名等	発表年月
Ontologizing Lexicon Access Functions based on a LMF-based Lexicon Taxonomy	Proc. LREC2008	2008 年 6 月
複合的な語彙資源アクセスサービスの実現基盤	言語処理学会第 15 回年次大会	2009 年 3 月