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

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　京都大学大学院情報学研究科研究員(現在，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, 2007a)．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いて，辞書言語資源を組み合わせて利用するために必要となる基盤技術の研究を進めた．

2 異文化間コミュニケーション支援と言語資源の連携

異文化間コミュニケーションという言葉の指す範囲は大変に広く，様々なタイプの異文化間コミュニケーションにおいて，どのような言語コミュニケーションの支援が必要・有用であるかは必ずしも明らかではない．そこで，本研究の当初計画においては，異文化間コミュニケーションの実態を調査し，その類型化を行うとともに，そこにおいて求められる言語コミュニケーション支援を明らかにしようとしたが，実態の調査には様々な困難が伴うことが明らかとなったので，計画を変更し，文献調査を中心に検討を進めた．

検討においてとくに参考としたのは，外国語教育の分野における学習者の辞書使用に関する議論の中で，単言語辞書(国語辞典や英英辞典など)と対訳辞書(和英辞典や英和辞典)を組み合わせることの有用性である(Hartman, 2005)，(Laufer et al, 2006)．これらの議論においては，多くの対訳辞書が単語の対照にとどまっており，意味概念や使用場面にかける情報にかけていることから，これらの情報を豊富に持つ単言語辞書と組み合わせて用いることの有用性が論じられている．例えば，日本語母語話者が英語でのコミュニケーションを行う場合，日本語の概念辞書により伝えたい概念を確認しながら，その意味概念になるべく近い英語単語を選べることが望ましい．このような辞書の連携は，外国語教育，あるいは，応用辞書学の分野では二言語化辞書(bilingualized dictionary)と呼ばれており，まさに本研究が対象とする言語基盤が目指す言語資源の組み合わせによって実現が可能となる．

そこで，本研究では，意味概念を扱っている代表的な辞書であり，入手が容易で，計算機上でも容易に利用できる辞書として WordNet(英語)(Fellbaum, 1998)，EDR 電子化辞書(日本語，英語)(EDR, 2003)，また，Web 上でサービスとして利用可能となっている人間用の辞書も検討の対象とし，これらの言語資源を連携させるための基盤技術の検討を進めた．

3 言語基盤と言語サービスオントロジー

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．言語基盤は，その目的の観点から，サービス指向の言語基盤(service-oriented infrastructure)と研究のための言語基盤(research infrastructure)に分類することができる．筆者がかかわっている NICT の言語グリッド (http://langrid.nict.go.jp) は前者の例であり，異文化コラボレーションを支援することを目的としている．一方，後者の例としては，欧州における CLARIN (Common Language Resources and Technology Infrastructure) (Calzolari, 2008)と呼ばれる e-humanity 分野の研究をターゲットとするプロジェクトがあげられる．これらの言語基盤の目的は大きく異なっている

が，いずれにおいても，その言語基盤の上で言語データ資源や言語処理ツールを組み合わせ，所望の言語サービス機能を得ようとすることは共通している．また，この共通の目的に関連し，言語資源の再利用性(re-usability)，相互運用性（interoperability）といった課題も共有している．

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．この場合，個々の構成要素の詳細を隠蔽しAPIを実装するラッパーが必要となる．たとえば，データ的な言語資源は，言語データへのアクセス機能を持つラッパーにより言語サービス化される．ここで，構成要素となる言語資源や言語処理機能には多様なタイプが想定されるため，APIはこれらのタイプに応じて設定することが必要となる．また，新たに開発された言語資源や言語処理機能をWebサービス化する場合，そのタイプに応じたAPIを選択または実装し，ラッパーを準備する必要がある．さらに将来的に，ゴール記述をもとに複合的な言語サービス（composite Web service）を自動構成する場合には,構成要素の機能や入出力に関して,共有された形式的な枠組みによって与えられた記述が必要である．言語サービスオントロジーはこのための基盤を与える．

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. 2008)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．言語サービスオントロジーは，OWL（Web Ontology Language）という言語を用いて記述されているため，国際標準の仕様をOWL言語で記述することになる．

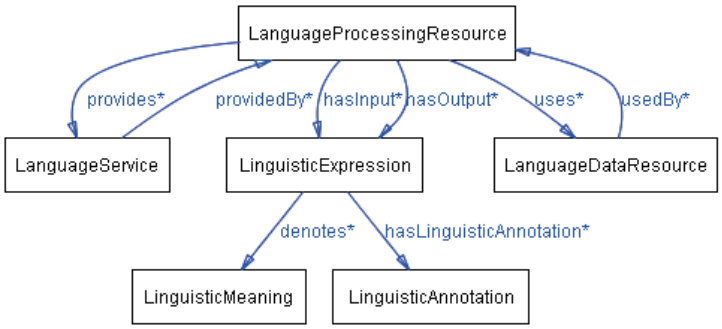

図 1: 言語サービスオントロジーの最上位階層

4 辞書言語資源のモデル化

辞書言語資源のサブオントロジーにおいては，代表的な辞書のタイプを定義し，それらの分類体系を提示する必要がある．後に詳述するように，辞書のタイプおよび分類体系は，おもにその辞書がどのようなタイプの辞書エントリを有するかによって定まる．ある辞書をモデル化することは，その辞書が持つ辞書エントリのタイプを記述することである．また，このような辞書のモデルにおいては，モデルを記述するための枠組みが必要となる．このような枠組みをメタモデルと呼ぶことがある(林, 2007b)．辞書のメタモデルに関しては，国際標準化機関であるISO（International Standardization Organization)において，LMF（Lexical Markup Framework）(Francopoulo et al. 2006)という標準化案が検討されてきており，間もなく最終的な国

際標準として制定される予定になっている．

LMF は，あらゆるタイプの辞書をモデル化するための枠組み(メタモデル)を規定することを目的としている．多様なタイプの辞書をモデル化するための包括的な枠組みを提示するにあたっては，モジュラーな構造をとっている．すわなち，LMF のモデルは，すべてのタイプの辞書に共通する規定である Core Model と代表的なタイプの辞書を規定するための Extensions から構成される．

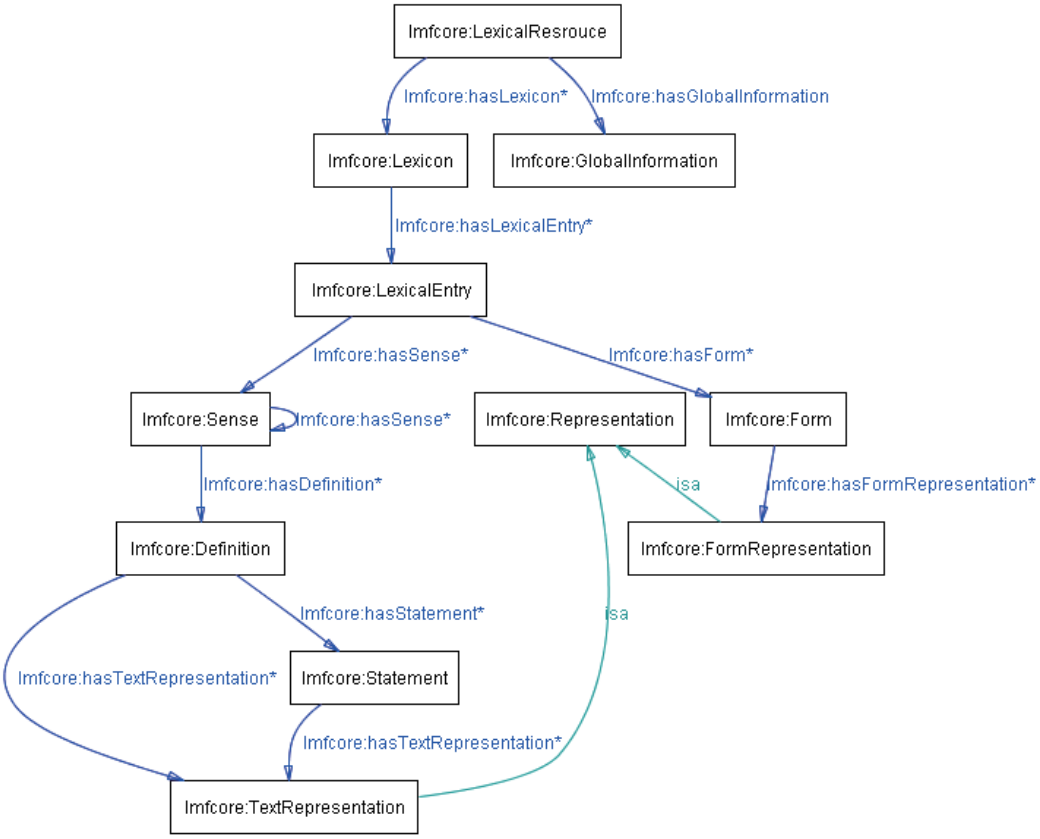

図 2: LMF Core Model のオントロジー的規定

図 2 に LMF 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．LMF の仕様は，UML (Unified Modeling Language) という図式表現により与えられているが，一定の変換の対応を定めることにより，そのオントロジー化を行うことができる．すなわち，UML における generalization は OWL においては subclass を用いて表現できる．また，aggregation は，**hasX** などの適当な名前の property を導入できることにより表現できる．LMF における各 Extension は，Core Model を拡張することにより定義されているが，オントロジー化においては必要なクラスをサブクラス化し，必要な属性や関係を付加していくことにより規定することができる．

図 3 に NLP Semantics に関する Extension の一部(特に語彙意味論的関係をカバーする範囲．統語論と意味論の間のリンクは範囲外)のオントロジー的規定を示す．ここで注意すべきことは，辞書のサブクラスは，それが持つ辞書エントリのクラスにより決定されるということである．図 3 において意味辞書(**lmf:Sem.Lexicon**)は，**lmf:Sem.LexicalEntry** というクラスにより規定される辞書エントリを有する辞書として定義されている．ここで，**lmf:Sem.LexicalEntry** とは，図にあるように，**SenseRelation**, **MonolingualExternalRef**, **SenseExample** といたプロパティを有すということにより規定されることに注意されたい．

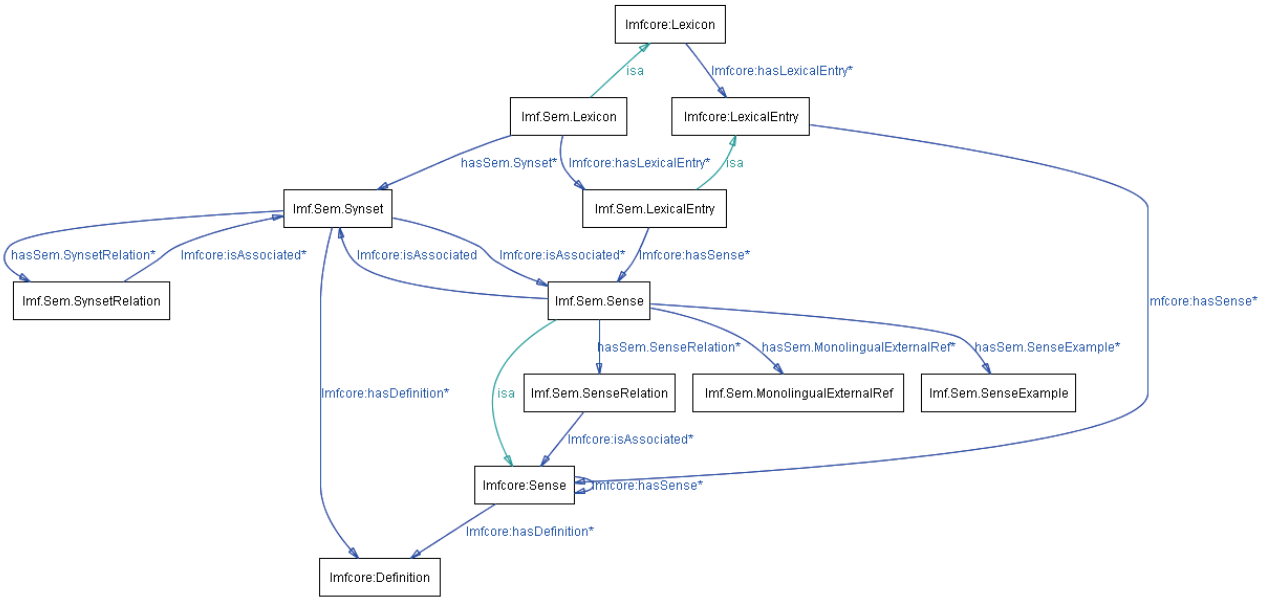

図 3: LMF NLP Semantics Extension (一部)のオントロジー的規定

図 1 に示した言語サービスオントロジーの最上位階層における言語データ資源クラス(**LanguageDataResource**)は，言語基盤において対象とする言語データ資源のさまざまなタイプに応じて詳細化される．たとえば，その直下では，辞書に関するクラス(**Lexicon**)とコーパスに関するクラス(**Corpus**)にサブクラス化され，それぞれがサブオントロジーを構成する．また，これらの言語データ資源にアクセスする言語処理資源もこれらのサブオントロジーに応じて詳細化される．図 4 にこの様子を示す．図 4 においては，言語処理資源(**LanguageProcessingResource**)の下位クラスである言語資源アクセッサ(LR_**Accessor**)が利用する言語データ資源のサブクラス化(ここでは **Corpus** と **Lexicon**)に応じてサブクラス化(**CorpusAccessor** と **LexiconAccessor**) されている．

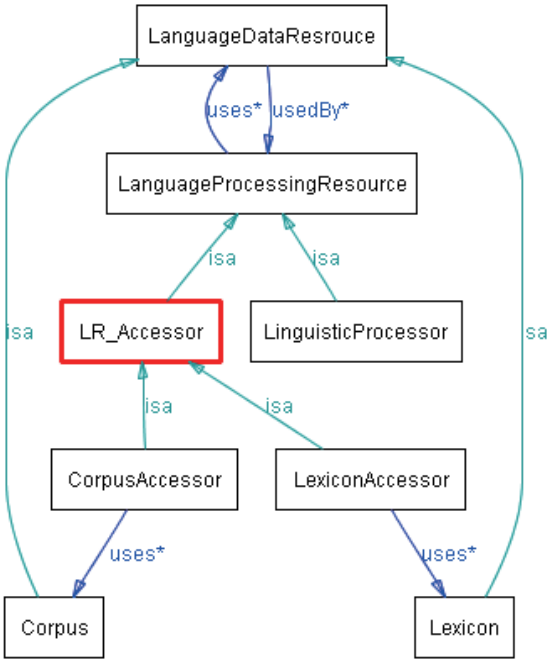

図 4: 言語データ資源と対応する言語処理資源のオントロジー的規定

ここで注意すべきことは，これらの言語データ資源のサブオントロジーをどのように詳細化していくかということは多分に対象とする言語基盤の目的や特性に依存する可能性があるということである．したがって，

特定の言語基盤の目的や特性に基づく辞書のサブオントロジーは，より多くのユーザに共有されている標準的な体系に立脚するような構造になっていることが望ましい．そこで，特に辞書に関するサブオントロジーについては，これをオントロジー化された LMF（以下，LMF オントロジー）と対応付ける．これにより，言語基盤におけるサブオントロジーが国際標準によって基盤付けられることになる．すなわち，LMF が国際標準として実効的である限り，それに立脚する辞書サブオントロジーは，LMF という国際標準がもたらすメリットを享受することができる．

図 4 における辞書クラス（**Lexicon**）は，たとえば以下のようなサブオントロジーとして規定される．まず，人間用の辞書（**DictionaryForHumanUse**）と言語処理用の辞書（**LexiconForNLP**）にサブクラス化される．前者の下位分類としては，いわゆる通常の機械可読形式辞書（Machine Readable Dictionary）（**MRD**）と専門用語集（**Terminology**）が設定されている．さらに MRD は単言語辞書（**MonolingualDictionary**）と対訳辞書（**BilingualDictionary**）に下位分類されている．このようなサブオントロジーは多分に恣意的であり，必ずしも多くの納得が得られる辞書学的に適正な構造となっているとは限らない．そこで，このようなサブオントロジーを LMF オントロジーと対応付けることにより，その基盤を与える．図 5 に上記の趣旨の辞書サブオントロジーを LMF オントロジーへ対応付けた例を示す．

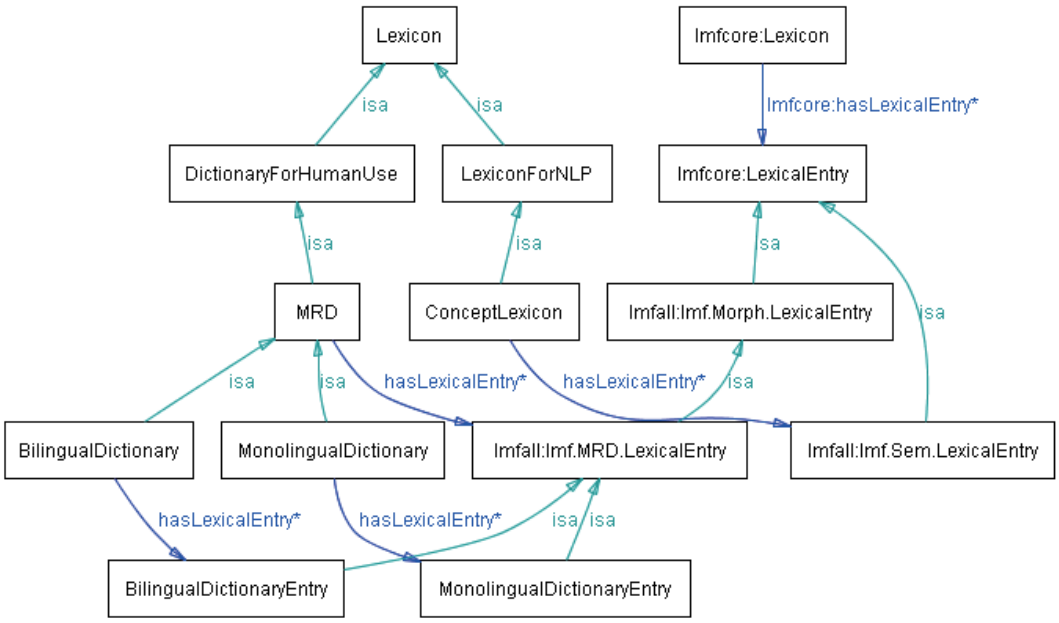

図 5 辞書クラスサブオントロジーの LMF オントロジーへの対応付け

LMF においては，機械可読な人間用の辞書に対する MRD extension を用意している．そこで，図 5 における MRD クラスは，LMF オントロジーにおいて，MRD extension で定義される辞書エントリ（**lmfall:lmf.MRD.LexicalEntry**）を持つような辞書であると定義する．さらに，その下位分類である単言語辞書や対訳辞書における辞書エントリを LMF オントロジーにおける辞書エントリクラスである **lmfall:lmf.MRD.LexicalEntry** の下位分類である辞書エントリを有する辞書であると定義する．このように，LMF における extension に準じて，オントロジー化された言い方でいえば，LMF オントロジーにおけるサブクラスに準じて，LMF オントロジーを詳細化する形で適切な辞書エントリのサブクラス化を行うことができれば，それをエントリとして持つものとして新たなタイプの辞書言語資源を規定していくことが可能となる．

5 複合辞書言語資源のモデル化

複数の原子的な（atomic）言語サービスを組み合わせることにより，利用者の目的に適合した複合的な（composite）言語サービスを構成することがサービス指向の言語基盤の目的の一つである．原子的なサービスの組み合わせを規定するのが，サービスワークフロー（service work flow）である．例えば，日本語から英語への機械翻訳と英語からドイツ語への機械翻訳を直列に接続することにより日本語からドイツ語への翻訳機能を（翻訳精度の可否は別として）実現することができる．また，特定の形態素解析の結果を受けて，構文解

析を実行することも可能である．

辞書言語資源へのアクセスにおいても同様に，特定の辞書へのアクセスを組み合わせることにより，複合的な辞書アクセス機能を実現することが考えられる．たとえば，英和辞書の基本的な機能は，英語の各単語に対する語義ごとに日本語の訳語を提示することであるが，ここでの語義と WordNet における語彙概念を対応させ，WordNet における synset に関する情報と日本語の訳語を同時に提示するような複合辞書サービスを考えることができる．このような複合的な辞書アクセス機能を実現することは，英和辞書と WordNet を組み合わせた仮想的な複合辞書を構成することと等しい．

このような複合辞書をオントロジー的に規定するためには，要素となる辞書の辞書モデルから，当該の複合辞書のオントロジー的規定を合成できることが必要となる．このことを実現するためにまず考えられる方法は，クラスの多重継承(multiple inheritance) を利用する方法である．すなわち，英和辞書が属する対訳辞書(**BilingualDictionary**)クラスと WordNet が属する概念辞書(**ConceptLexicon**)クラスの双方を親クラスとして持つクラス(**BilingualConceptLexicon**)を定義する．図 6 にこの様子を示す．対訳辞書クラスは **MRD.LexicalEntry** クラスで規定される辞書エントリを持つものとして定義され，概念辞書クラスは，**Sem.LexicalEntry** クラスで規定される辞書エントリを持つ．

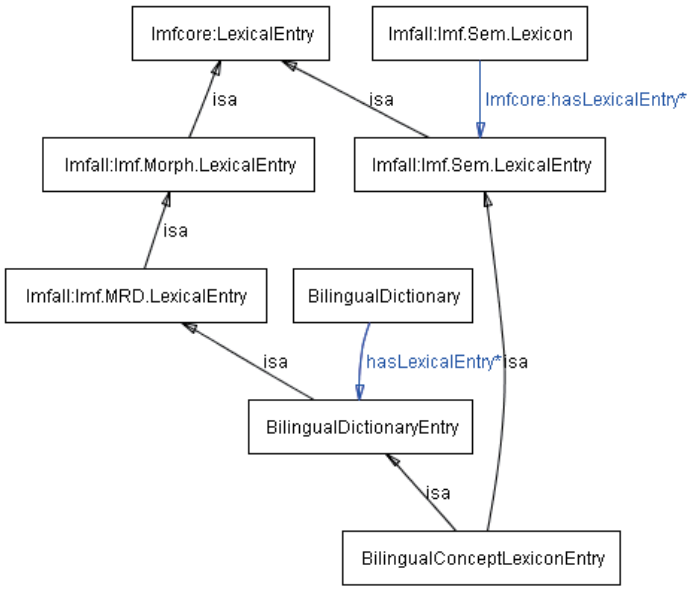

図 6: **BilingualConceptLeixonEntry** のオントロジー的規定

6 関連研究・プロジェクト

Princeton 大学による英語の語彙データベースである WordNet(Fellbaum, 1998)は，自由に利用できる大規模な言語資源として大きなインパクトを与えた．ヨーロッパにおいては，EuroWordNet (Vossen, 1998)と呼ばれるプロジェクトが実施された．EuroWordNet においては，Princeton 大学の WordNet の構造に従って構築された各国語の wordnet を ILI(Inter-Lingual Index) と呼ぶ構造により相互に接続することにより，多言語の概念辞書資源(あるいは言語的オントロジー)を実現する方法を示した．

さらに近年，EWN の考え方をさらに発展させ，世界各国語の wordnet を連携させることを目的とする Global WordNet Grid (Vossen and Fellbaum, 2007) が提唱されている。Global WordNet Grid においては、各 wordnet を結び付けるバックボーンの一部として SUMO (Suggested Upper Merged Ontology) (Niles, 2003) と呼ばれる上位オントロジーを用いることを想定している．また，本報告でも述べた LMF を用いて各 wordnet の連携をモデル化する試み (Vossen et al., 2008) も開始されている．

7 おわりに

本報告では，とくに辞書言語資源の連携を実現する際に必要となる辞書言語資源のモデル化と国際標準に基づくオントロジー化の検討状況について報告した．この結果は，NICT における言語グリッド基盤における

言語サービスオントロジーへと反映されている．本研究は 2008 年度も継続が決定しており，2008 年度においては，実際に EDR 電子化辞書，WordNet を連携する枠組みを LMF に基づいて記述し，さらに，Web サービスのプロトタイプを実装することにより，技術の確認とさらなる研究課題の抽出を行う予定である．また，このために必要な辞書エントリ間の対応付け手法とその詳細な表現モデルについても実証的に研究を進めていく予定である．

【参考文献】

1. 林 良彦. (2007a). 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. (2007b). 再利用・相互運用可能な言語資源の記述とモデル化の枠組み. 電子情報通信学会論文誌 D, Vol.J90-D, No.12, pp.3114-3130.
3. 林 良彦. (2008). 言語の意味・概念への計算機科学からのアプローチ.『言語文化学への招待（木村・金崎編)』大阪大学出版会, pp.221-234.
4. Calzolari, N. (2008). Approaches towards a “Lexical Web”: The Role of Interoperability. *Proc. of ICGL2008*, pp.34-42 (invited talk).
5. EDR. (2003) EDR Electronic Dictionary Technical Guide. http://www2.nict.go.jp/kk/e416/EDR
6. Fellbaum, C. (Eds.). (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.
7. 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.
8. Laufer, B., and Levitzky-Aviad, T. 2006. Examining the Effectiveness of 'Bilingual Dictionary Plus' - A Dictionary for Production in a Foreign Language. *International Journal of Lexicography*, Vol.19, No.2, pp.135-155.
9. Hartman, R.K.K. 2005. Pure or Hybrid? The Development of Mixed Dictionary Genres. *Linguistics and Literature*, Vol.3, No.2, pp.193-208.
10. Hayashi, Y. (2007). A Linguistic Service Ontology for Language Infrastructures. In: *Proc. of ACL (poster proceedings)*. pp.145-148.
11. Hayashi, Y., Declerck, T., Buitelaar, P., and Monachini, M. (2008). Ontologies for a Global Language Infrastructure. In: *Proc. of ICGL2008*, pp.105-112.
12. Niles, L. and Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In: *Proc. of IKE03*, pp.23-26.
13. Vossen, P. (Eds.) (1998). E*uroWordNet: A Multilingual Database with Lexical Semantic Network.* Kluwer Academic Publishers.
14. Vossen, P. and Fellbaum, C. (2007). Connecting the Universal to the Specific: Towards the Global Grid. In: Ishida, T. et al. (Eds.) *Intercultural Collaboration*, LNCS4568, Springer., pp.1-16.
15. Vossen, P. et al. (2008). KYOTO: a System for Mining, Structuring and Distributing Knowledge across Languages and Cultures. In: *Proc. of LREC2008*, pp.373-380.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
セマンティック Web と言語技術・言語資源	情報処理	2007 年 8 月
再利用・相互運用可能な言語資源の記述とモデル化の枠組み	電子情報通信学会論文誌 D	2007 年 12 月
言語の意味・概念への計算機科学からのアプローチ	『言語文化学への招待, (木村・金崎編)』 (大阪大学出版会)	2008 年 3 月
A Linguistic Service Ontology for Language Infrastructures	The 45th Annual Meeting of the Association for Computational Linguistics (ACL2007). (Poster proceedings)	2008 年 6 月
Ontologies for a Language Infrastructure	The First International Conference on Interoperability for Language Resources (ICGL2008)	2008 年 1 月