言語サービスオントロジーの開発における国際標準案の適用

Incorporating Relevant International Standards in Developing a Language Service Ontology

林 良彦[1], 楢和千春[2], Thierry Declerck[3], Paul Buitelaar[3,4], Monica Monachini[5]

[1]大阪大学大学院言語文化研究科 *Graduate School of Language and Culture, Osaka University*

[2]京都大学大学院情報学研究科 *Graduate School of Informatics, Kyoto University*

[3]*DFKI GmbH, Language Technology Lab, Germany*

[4]*DFKI GmbH, Competence Center Semantic Web, Germany*

[5]*Instituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche, Italy*

1 はじめに

辞書やコーパスなどの言語資源，言語解析や機械翻訳などの言語処理機能が開発され，利用可能となっている．また，さまざまなコミュニティにおいて，独自の用語集や対訳集などの言語資源が構築される例も増加している．このような広義の言語資源を目的に応じて組み合わせることにより，言語に関する有用なサービス(言語サービス)がWeb上で利用可能になると期待され，言語サービスの作成，運用，利用を可能とする言語基盤の実現が求められている．

Web上のオープンな環境において，上記のような言語基盤を実現するためには，サービスの基本要素を規定するためのオントロジー的な共有基盤(言語サービスオントロジー)が必要となる．筆者らは，異文化コラボレーションを対象とする言語基盤である言語グリッド(Ishida, 2006)をターゲットとして，言語サービスオントロジーの上位レベルの構成を提案してきた(林，楢和 2007)．本稿ではさらに，関連する言語資源に関する国際標準案を適切に利用することの必要性を論じ，実際に言語的注釈や辞書のメタモデルに関する国際標準案を言語サービスオントロジーに取り込む方法論について議論する．

2 言語基盤における言語サービスオントロジーの役割

2.1 言語資源の相互運用性：言語基盤における課題

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．上述の言語グリッド[1]は，異文化コラボレーションを支援することを目的とするサービス指向の言語基盤である．一方，欧州においては，CLARIN (Common Language Resources and Technology Infrastructure)[2]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立ち上がっている．

これらのプロジェクトの目的は大きく異なっているが，いずれにおいても，言語基盤の上で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(林，2007)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．

2.2 言語サービスオントロジー

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．

さて，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．言

[1] http://langrid.nict.go.jp/

[2] http://www.clarin.eu/

語サービスオントロジーの開発においては Protégé[3]と呼ばれるツールを用い，OWL(Web Ontology Language)による記述を進めている．

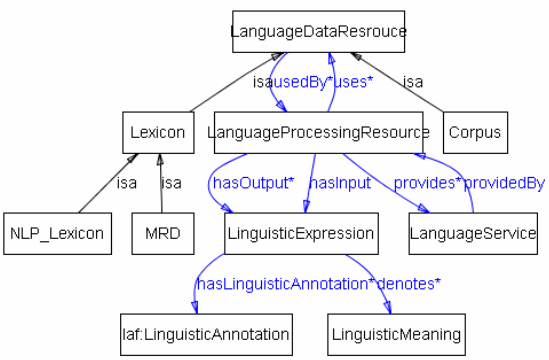

図 1：言語サービスオントロジーの最上位階層

図 1[4]にわれわれが提案する言語サービスオントロジー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(**laf: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．たとえば，言語データ資源は，辞書(**Lexicon**)やコーパス(**Corpus**)クラスに下位分類され，さらに辞書クラスは，人間用の辞書(**MRD**：Machine Readable Dictionary)とコンピュータ処理用辞書(**NLP_Lexicon**)に細分類される(Hayashi et al. 2008)．

3 言語資源に関する国際標準案

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．

言語資源に関しては，国際標準化機構 ISO における TC37/SC4[5]が関連する標準化案の策定を行っている(Declerck, et al. 2008; 林, 2007)．**図 1** に示した最上位階層に関連しては，言語的注釈および辞書に関する標準化が進行している．また，この双方に関連して，言語的な属性・属性値に関する標準化が検討されている．

- LAF (Linguistic Annotation Framework)：言語データ (primary data)に関する言語的注釈に関する一般的な枠組みである．言語データに対する注釈を言語データと分離することにより，多重の言語注釈を与えることを可能としている．言語的な注釈の内容自体は，素性構造により表現される．形態統語論的な注釈のための枠組み MAF (Morphosyntactic Annotation Framework)，統語論的な注釈のための枠組み SynAF (Syntactic Annotation Framework) は，LAF の特殊化されたものである．
- LMF (Lexical Markup Framework) (Francopoulo, et al. 2006)：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．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．仕様[6]は UML(Unified Modeling Language)を用いて示されている．
- DCR (Data Category Registry)：言語的注釈における言語的属性(データカテゴリ)とそれらが取りうる atomic な属性値を規定するための枠組みである．

これらの関連する国際標準案を言語サービスオントロジーに取り込むために，国際標準案の *オントロジー化* (ontologization)を行う．ここで，オントロジー化とは仕様に対応する OWL 記述を与えることを意味する．

4 LAF のオントロジー化

図 2 に言語的注釈に関するオントロジーの上位レベルを示す．

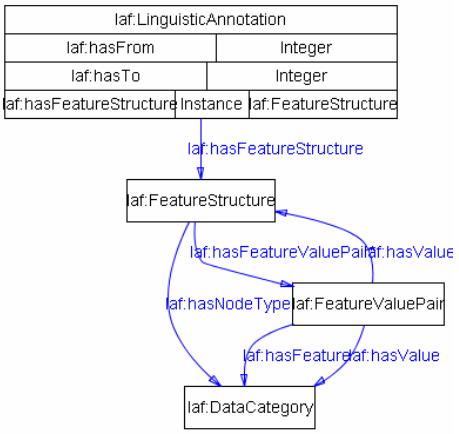

図 2：LAF のオントロジー的規定

[3] http://protege.standord.edu/
[4] 以下同様の図は Protégé の Ontoviz プラグインによる．
[5] http://www.tc37sc4.org/
[6] 仕様書は http://www.lexicalmarkupframework.org/ より入手可能．

hasFrom, **hasTo** という属性により注釈を与える言語データの区間(span)を指定する．また，言語的注釈の内容を素性構造(**laf:Feature Structure**)により与える．この上位レベルを詳細化することにより，MAF や SynAF に対応する言語レベルの言語的注釈の構造をオントロジー的に規定することができる(Hayashi, et al. 2008)．

ここで，言語的属性とその atomic な属性値については，データカテゴリに関するクラス(**laf:DataCategory**)により定義される．このクラスのサブオントロジーは，ISO DCR（あるいは同様の言語的属性オントロジー）のオントロジー記述を import することにより規定される．

本稿ではその詳細を述べることはできないが，言語処理資源，とくに，対象の言語データに関してある言語レベルの言語的注釈を加える言語解析器は，その入力として想定する言語的注釈のクラス，その出力とする言語的注釈のクラスにより分類され，そのタクソノミーを構成する(林，楢和，2007)．すなわち，言語処理資源に関するサブオントロジーを基盤付けるためにも言語的注釈のオントロジー化は重要な役割を果たす．さらに，いわゆるタグ付きコーパスの内容は言語的注釈そのものであり，言語的注釈に関するサブオントロジーは，コーパス言語データ資源に関するサブオントロジーの基盤付けにも寄与する．

5 LMF に基づく辞書サブオントロジー

5.1 LMF のオントロジー化

前述のように LMF の仕様は UML のクラス図により与えられているが，そのオントロジー化は容易である．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．**図 3** に LMF Core Model のオントロジー的規定を示す．また，LMF の各 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も同様に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．

5.2 辞書サブオントロジー

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 **Lexicon** クラスは，言語基盤において対象とするさまざまな辞書のタイプに応じて，サブオントロジーとして詳細化される．このサブオントロジーは，オントロジー化された LMF(以下，LMF オントロジー)と対応付けることにより，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 に言語サービスオントロジーにおける，**Lexicon** クラスを最上位クラスとする辞書サブオントロジー詳細化の概念を示す．**図 4** に明らかなように，言語サービスオントロジーにおける **Lexicon** クラスは，LMF オントロジーにおける **lmf:LexicalEntry** クラスによって規定される辞書エントリを持つ(**hasLexicalEntry**)言語資源として定義される．また，Lexicon クラスの下位分類として位置づけられる各種の辞書は，それに応じたタイプの辞書エントリを持つ言語資源として定義される．たとえば，対訳辞書(**BilingualDictionary**)は，人間用の機械可読辞書(**MRD**)の下位分類として位置づけられ，特定のクラス(**BilingualLexicalEntry**)により規定される辞書エントリを持つ言語資源として定義される．ここで，**BilinguaLexicanEntry** は，LMF オントロジーにおいて規定される **lmf:mrd.LexicalEntry** クラスの下位クラスとして規定され，このクラスはさらに，**lmf:morph.LexicalEntry** のサブクラスとなっている．このような構成は，LMF の Extensions により規定されているものである．LMF における Extension 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．

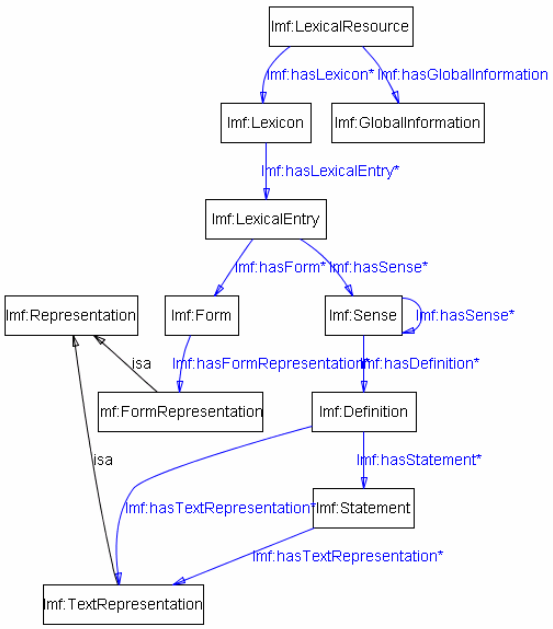

図 3: LMF Core Model のオントロジー的規定

5.3 辞書アクセス機能のオントロジー

すでに述べたように，言語データ資源はそのアクセス機能によってラップされ言語サービス化される．このような言語データ資源へのアクセス機能は，言語処理資源の下位分類をなし，利用する言語データ資源のタイプや入出力データのタイプによってさらに詳細化される．

辞書アクセス機能は，言語表現クラスの特殊なサブクラ

スである辞書アクセスクエリ(**LexiconAccessQuery**)を入力とし,同じく言語表現クラスのサブクラスである辞書アクセス結果(**LexiconAccessResult**)を出力するものとして定義されるが,ここで,辞書アクセス結果はクエリに与えられた言語表現に対する辞書的意味を指示する(**denotes**)ものとして規定される(林,楢和,2007).

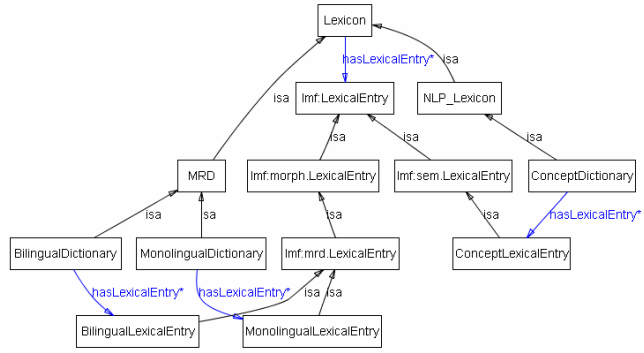

図 4:辞書に関するオントロジー的規定

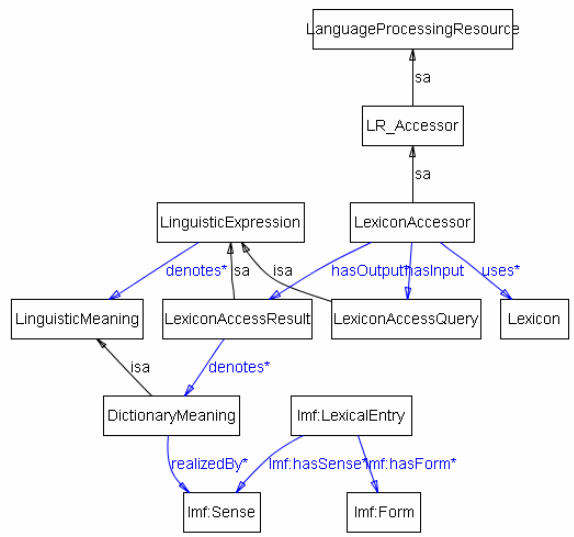

図 5:辞書アクセス機能に関するオントロジー的規定

図 5 に辞書アクセス機能に関するサブオントロジーを示す.辞書的意味(**DictionaryMeaning**)として,辞書エントリに記載されているどのような情報内容が返されるかの詳細はアクセス対象の辞書の特性により定まるが,**図 5** に示す上位レベルでは,LMF オントロジーにおける **lmf:Sense** という一般的なクラスによって規定される情報内容によって定まる(**realizedBy**)としている.

なお,辞書アクセスが成功した場合,アクセスクエリに指定された言語表現と当該の辞書エントリにおける見出し語(**lmf:Form** クラスで規定される)は,クエリで指定された検索条件に応じた関係が存在するはずであるが,「処理が成功した場合」というような動的な制約は,現状のオントロジーの範囲外であり,今後の課題のひとつである.

6 おわりに

本稿では,言語注釈(LAF)と辞書のメタモデル(LMF)に関する国際標準案の枠組みをオントロジー化し,言語サービスオントロジーに取り込む方法論について議論した.言語サービスオントロジーは,可能な限り関連する国際標準案に適合していることがその受容度を高めるために重要であり,また,言語グリッド以外の他の言語基盤との間の相互連携を図る際にも有用である.今後は,実際の言語資源や言語処理ツール・システムを記述していくことにより,言語サービスオントロジーの詳細化や拡張を行っていく.

なお,ISO 以外で注目すべき標準化指向のアクティビティとして,言語処理プラットフォーム UIMA(竹内ほか,2007)がある.とくに,言語データの型を規定するために標準化された Sharable Type System を定めようという動き(Kano, et al. 2008)は本研究と関係が深い.また,言語データの内容レベルでの相互運用性に関連しては,DCR のオントロジー化や GOLD (Farrar and Langendoen, 2003) のような言語的オントロジーの方向性について注目していく必要がある.

参考文献

竹内広宣, ほか. 2007. UIMA: 非構造情報処理アーキテクチャ. 人工知能学会誌, Vol.22, No.6, pp.806-813.

林 良彦, 楢和千春. 2007. 言語サービス記述のための上位オントロジーの提案. NLP2007, B5-3, pp.1101-1104.

林 良彦. 2007. 再利用・相互運用可能な言語資源の記述とモデル化. 電子情報通信学会論文誌 D, Vol.J90-D, No.12, pp.3114-3130.

Thierry Declerck, et al. 2008. Interoperable Language Resources. *Sprache und Datenverarbeitung (International Journal for Language Data Processing),* forthcoming.

Scott Farrar, and Terry Langendoen. 2003. A Linguistic Ontology for the Semantic Web. *Glot International*, Vol.7, pp.97-100.

Gil Francopoulo, et al. 2006. Lexical Markup Framework (LMF). *Proc. of LREC2006*, pp.233-236.

Yoshihiko Hayashi, et al. 2008. Ontologies for a Global Language Infrastructure. *Proc. of ICGL2008.* pp.105-112.

Toru Ishida. 2006. Language Grid: An Infrastructure for Intercultural Collaboration. *Proc. of SAINT2006*, pp.96-100.

Yoshinobu Kano, et al. 2008. Sharable Type System Design for Tool Interoperability and Combinatorial Comparison. *Proc. of ICGL2008*, pp.122-129.