人工知能学会研究会 資料 SIG-SWO-040-04

日本語 WordNet の語彙拡充のための文法オントロジの作成と活用

Grammer Ontology for Expanding Japanese WordNet Vocabulary

小林賢司 [1*]　鵜飼孝典 [1]　井形伸之 [1]　西野文人 [1]

[1] (株) 富士通研究所
[1] Fujitsu Laboratories Ltd.

Abstract: It requires a larger dictionary to make an application to interact more naturally with the users. Japanese WordNet is one of the free dictionaries, which includes thesaurus. It has RDF formed dataset that links to other resouces such as DBpedia, so it is useful for our supposing application. However the WordNet is insufficient because of the small volume of vacabulary, lack of parts of speech and derivative relations, and so on. Japanese Wiktionary, which is another free dictionary, has the parts of speech and the derivative relations. It is expected that more vocabularies can be extracted than from the WordNet. We have built a glammar ontology based on the structure of the Wiktionary to express of missing informations in the WordNet ontology. The volume of vocabularies is expanded to 1.16 times compare with Japanese WordNet with the ontology. 33,909 links are added about parts of speech, 317 links about derivative relation, and so on. The application can rephase over the part of speech using the extended dictionary.

1　はじめに

人工知能分野の活発化に伴い，対話技術や質問応答技術の発展が一層期待されている．その中で我々は，アプリケーションがテキストあるいは音声を入力とした自然言語から，正しく意味を解釈し，自然な回答を出力できることを目指している．そのためには，語句の意味や概念だけでなく，活用や用法，読み，発音などを含んだ，構造化された機械判読可能な辞書データが必要となる．例えば,「目的地まで歩きたい」という入力に対し,「徒歩なら 10 分かかります」と返答するには，動詞 “歩く” の連用形が “歩き” であり，それを名詞化した “歩き” という派生語と “徒歩” が同じ概念である，という知識を保持する必要がある．

現在，公開されている代表的な日本語辞書データとしては，日本語 WordNet[1] や日本語 Wiktionary が挙げられる．日本語 WordNet は，シソーラスであり，語句の意味，同義語，上位／下位概念などがまとめられている．また，データは RDF(Resource Description Framework) で提供され，DBpedia など他のリソースともリンク付けられている．RDF は, Web において情報を記述するグラフベースのデータモデルに基づいた形式的言語であり, Web の情報をソフトウェアによる自動処理などに使われることを想定して作られている [2]. そのため，上記アプリケーションを想定する本研究にとって，日本語 WordNet を利用することは都合が良い．しかし，語彙量は十分ではなく，品詞や活用形の情報が貧弱，派生語の登録有無が曖昧，などの問題がある．

一方，日本語 Wiktionary は，日本語 WordNet と異なり，RDF 化されておらず，概念体系について乏しいが，品詞や活用形，漢字・読み表記関係，訳語，発音，語源など，日本語 WordNet では不足している情報を多く持つ．また，日本語 WordNet が登録していない語句も持っており，更には活用による派生語や，漢字・読み表記関係まで考慮すると，日本語 WordNet 以上の語彙量を抽出できる可能性がある．

本研究は，日本語 Wiktionary のデータを RDF 化し，日本語 WordNet のデータ（オントロジ）と統合することによって，日本語 WordNet で不足している語彙を拡充することを目的とする. 本論では，以下について述べる．

- 日本語 Wiktionary から抽出したデータを元に，品詞・活用形，派生語関係，表記関係について RDF 化した “文法オントロジ” を作成する．
- 文法オントロジを活用し，日本語 WordNet オントロジにおいて欠落した語句や品詞・活用形，派生語関係，表記関係を補完する．

上記を行った結果，語彙量としては，日本語 Wik-

* 連絡先:(株) 富士通研究所
〒211-8588 神奈川県川崎市中原区上小田中 4-1-1
E-mail: kobayashi_kenji@jp.fujitsu.com

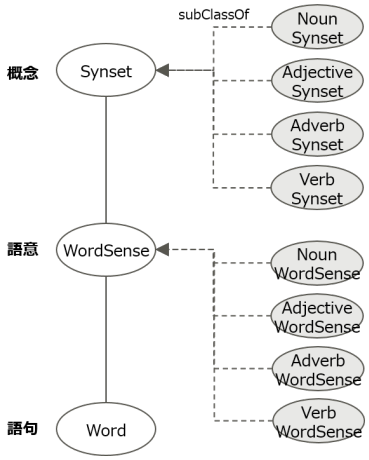

図 1　WordNet オントロジのクラス公理概要

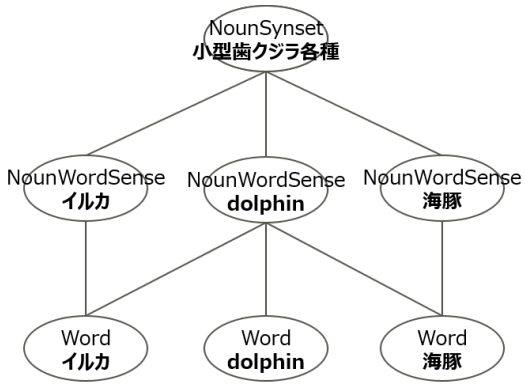

図 2　WordNet オントロジのインスタンス例

tionary が持つ 31,302 語に対して，派生語や表記関係から新規に 19,355 語を追加できた．この内，18,413 語は日本語 WordNet が持たない語句であり，全体の語彙量は約 1.16 倍となった．また，品詞・活用形については 33, 909 個，派生語関係については 317 個，漢字・読み表記関係については 28, 032 個のリンクを付与できた．これにより，派生語関係から自然な言い回しにアレンジしたり，テキストなら漢字，音声なら読みといった対応もしやすくなることが期待される．

本稿の構成は次の通りである．第 2 節では，日本語 WordNet とその問題点について述べる．第 3 節では，日本語 Wiktionary について述べる．第 4 節では，本稿の課題について述べる．第 5 節では，作成した文法オントロジについて述べ，第 6 節で，その活用について述べる．第 7 節では，活用により，拡充した語彙についての評価結果と考察について述べる．第 8 節では，関連研究について述べ，最後の第 9 節でまとめと今後の課題について述べる．

2　日本語 WordNet

日本語 WordNet は，プリンストン大学で開発された Princeton WordNet をベースとして，日本語向けに開発された WordNet である．WordNet は RDF データも公開されており，日本語 WordNet もこれに則して記述している．語彙量としては，約 9 万語が収録されている．

WordNet オントロジのクラス公理概要を図 1 に示す．WordNet オントロジは，概念 (Synset)，語意 (WordSense)，語句 (Word) の 3 クラスから成る．概念は，同義語となる語句の集合を表すクラスである．概念間では上位・下位，類似などの関係も定義される．語意は，語句の 1 つの意味を示し，概念と語句を紐付けるクラスである．語句は，少なくとも 1 つの意味を持つ表記を示すクラスである．また，概念上の品詞として，名詞・動詞・形容詞・副詞の 4 種に分類されており，概念または語意のサブクラスとして表現される．各国語固有の品詞が登録される際は，その 4 種いずれかに分類される．例えば，日本語における形容動詞や連体形は，形容詞に分類される．

インスタンス例を図 2 に示す．“イルカ”，“dolphin”，“海豚” という語句は，“小型歯クジラ各種” という共通の名詞概念を表す語意を持つ．なお，この例では，“dolphin” の 1 つの語意を表す語句として，“イルカ” や “海豚” も含まれる．

2.1　問題点

自然な対話や質問応答を実現するに当たって，日本語 WordNet を使用するには以下の問題点がある．

語彙量が少ない

日本語の一般的な辞書における収録語彙量は 20 万語上あることから，より多くの語が登録されることが望まれる．例えば，岩波書店発行の「広辞苑第六版」においては，約 24 万語が収録されている．

品詞・活用形の情報が不足している

自然言語の解析や生成を行うには，言語に合わせた，形容動詞／助動詞など詳細な品詞情報や，形容詞や動詞など用言に分類される品詞の活用形情報が必要となる．しかし品詞については，概念品詞の 4 種に留まり，活用形については分類すらされていない．

派生語の登録が曖昧となっている

WordNet では，概念的な品詞が異なる場合は，派

生語であっても区別して登録される．例えば，形容詞 “美しい” と，その副詞的用法となる “美しく” は，両方とも登録される．しかし，実際には，派生語が揃って登録されていない語句がある．

漢字・読み表記関係がない

日本語は，ある語句に対して，主に漢字や平仮名の組合せで異なる表記を行えるため，表記が異なっても検索できることが望ましい．しかし，日本語 WordNet では，漢字・読み表記の関係にある語句は，同じ概念の語句として登録されているが，両者の関係は特に定義されていない．

以上により，日本語 WordNet の語句にヒットするように，アプリケーション側で意識しなくてはならない．例えば，形容動詞 “綺麗” の副詞的語意である “綺麗に” の概念を知りたいとき，“綺麗” で引くべきか，あるいは，かな表記の “きれいに” で引くべきか，判断がつきにくいため，いずれも登録されていることが望ましい．

3 日本語 Wiktionary

日本語 Wiktionary は，Wiki を使った参加編集型の辞書サービスである Wiktionary の日本語版である．1 語句に対して 1 ページで，活用や用法，漢字・読み表記関係，訳語，発音，語源など，日本語 WordNet では不足している語句のより詳細な情報を収録している．また，WordNet ほどではないが，上位語や下位語，類義語など関連語も収録しており，その面でも今後活用できると考えられる．収録される日本語は約 3 万語 (ページ) であるため，日本語 WordNet と比較すると見劣りするが，ページ内には前述した関連語や漢字・読み表記関係など，他の語句の情報も多く含まれているため，潜在的な語彙量は豊富と言える．

4 課題

前述した通り，日本語 Wiktionary は，活用後の語句や，漢字・読み表記関係まで考慮すると，日本語 WordNet 以上の語彙量を持っている可能性がある．そのため，本研究では日本語 Wiktionary のデータを使用して，日本語 WordNet の語彙拡充を狙い，以下を行う．

文法オントロジの作成

語句の品詞・活用形，派生語関係，漢字・読み表記関係を表すクラス・プロパティ公理を定義し，日本語 Wiktionary から抽出したデータを当てはめた文法オントロジを作成する．品詞・活用形については，学校文法に習った表現とし，また語意クラスが示すのは概念的な品詞であり異なるため，新規に語句タイプを示すクラスを定義した．派生語関係は，文法的に派生語の生成が可能な関係を示すプロパティ，漢字・読み表記関係は，語意との関係を考慮した表記関係を示すプロパティをそれぞれ定義した．

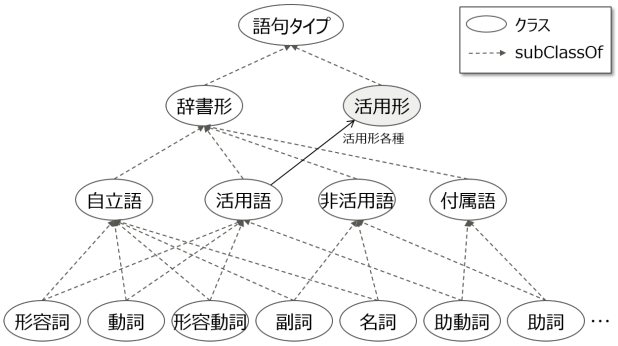

図 3　日本語文法上の品詞の表現

文法オントロジの活用

文法オントロジを日本語 WordNet オントロジに活用し，派生語関係，漢字・読み表記関係，及びそれらの関係にある語句を拡充する．文法オントロジは，語句の派生語関係および漢字・読み表記関係を持つが，語意とのリンクは日本語 Wiktionary が収録する語句に限られる．一方，日本語 WordNet は，前述の通り，派生語関係や漢字・読み表記関係を持たない．よって，文法オントロジを活用することによって，日本語 WordNet オントロジの語意に対してもこれらの関係をリンク付ける．

5 文法オントロジの作成

本節では，作成する文法オントロジについて述べる．文法オントロジは，品詞・活用形，派生語関係，漢字・読み表記関係を表現する．

5.1 品詞・活用形の表現

日本語文法上の品詞・活用形の表現について述べる．

日本語 Wiktionary では，基本的に学校文法を採用しているため，それに習った表現となるように，図 3 のようにした．WordSense クラスが示すのは概念的な品詞であり異なるため，新規に語句タイプを示すクラス (WordType クラス) を定義した．まず WordType クラスを，辞書形と活用形に 2 分した．辞書形は，活用されていない一般的な辞書に記載されている語形であり，基本

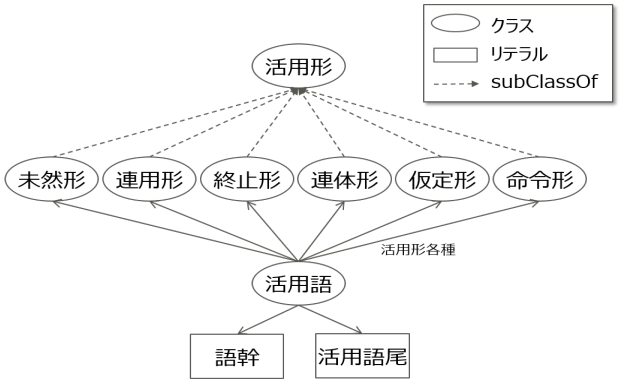

図 4　活用形の表現

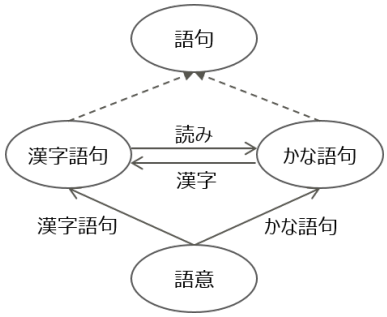

図 5　漢字・読み表記関係の表現

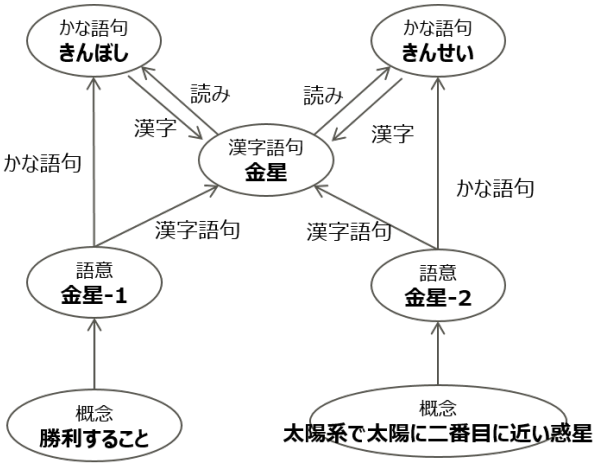

図 6　漢字・読み表記関係の具象化例

的には日本語 Wiktionary に登録されている語形である．学校文法においては，いずれかの品詞は，活用語または非活用語，および自立語または付属語に属するため，それに従ったクラス構造とする．また動詞の場合は，活用形を把握するために，五段活用や下一段活用など，活用形で更に分類する．

活用形は，活用語における活用された語形である．学校文法に習って未然形・連用形・終止形，連体形，仮定形，命令形を図 4 のように分類し，辞書形と，それぞれの形変化のプロパティとリンクする．具象化する際には，日本語 Wiktionry からは，品詞と活用形，語幹，活用語尾を取得できるため，まず辞書形を登録し，次に各活用形を生成し，辞書形とリンクを付けた上で登録する．

以上により，ある語句の品詞・活用形，および活用語の実際の語形を把握でき，自然言語文の解析または生成に使える他，後述する派生語関係の表現に対応できる．

5.2　派生語関係の表現

派生語を増やすために，文法的に異なる概念を派生することが出来る関係を定義する．文法的な派生語の例を表 1 に示す．例えば，動詞 “歩く” の連用形 “歩き” は，名詞として扱えることを示す．このような派生語関係毎に，プロパティを定義しておく．

具象化する際は，domain に合致する品詞・活用形の語句が日本語 Wiktionary から抽出された場合に，その派生方法に従って，派生語を生成することになる．これにより，日本語 Wiktionary のある 1 つの語句に対して，品詞を跨った派生語まで機械的に生成することが可能となる．

5.3　漢字・読み表記関係の表現

日本語 Wiktionary では，ある程度の漢字・読み表記関係を抽出可能である．また，同じ漢字でも複数の読みを持ち，更には読み毎に意味が異なる場合がある．逆に，同じ読みでも複数の漢字を持ち，漢字毎に意味が異なることが多い．よって，漢字・読み表記は意味と繋がる必要があるため，WordNet の語句クラスを継承し，語意クラスと紐付けることによって対応する (図 5)．図 6 は漢字，読み表記の具象化例である．この例における “金星” は，“きんせい” と “きんぼし” の複数の読みを持ち，また読み毎に意味が異なる．よって，“金星” は漢字語句クラス，“きんせい” と “きんぼし” は, かな語句クラスとし，双方を漢字または平仮名関係のプロパティでリンクする．また，各々の Word クラスは，その語意を示す WordSense クラスとリンクする．これにより，漢字・読み表記と語意を正しく表現することが可能となる．

表 1　文法的な派生語の例

派生元品詞 (domain)	派生方法	派生語品詞 (range)	例
動詞	連用形	名詞	歩く → 歩き
形容詞	連用形	副詞	美しい → 美しく
形容詞	語幹＋さ	名詞	美しい → 美しさ

6　文法オントロジの活用

文法オントロジと日本語 WordNet オントロジを照らし合わせ，品詞・活用形や派生語関係，表記関係の補完を行う．

品詞・活用形については，双方のオントロジが持つ語句と，その語句を持つ語意の概念的品詞が一致すれば，リンクを付与する．派生語関係についても同様に，派生語関係にある語句を持つ語意が存在し，その語意の概念的品詞が一致すれば，その語意間にリンクを付与する．例えば，文法オントロジにおいて，動詞 “走る” の活用形 (連用形名詞) は “走り” であるとする．ここで，“走る” の動詞的語意と “走り” の名詞的語意の表記を確認し，一致していれば，双方の語意をリンク付けする．これによって，ある語意に対して，別の品詞に言い換えた語意へと辿ることが可能となる．表記関係の補完については，日本語 WordNet オントロジにおける同一概念に属す語意の語句が，文法オントロジが持つ漢字・読み表記関係にある語句に一致するか，または，漢字語句を持つ語意に対して，読み表記となるかな語句とリンク付けする．

7　評価

文法オントロジを活用することによって，日本語 WordNet オントロジに追加可能な日本語の語彙量や，派生語関係，表記関係について評価した．日本語 Wiktionary の評価対象としては，日本語の意味を持つ語句に限定した．語彙量については，語句が重複していなければ，追加可能と判断した．品詞・活用形情報や漢字・読み表記関係を取得した語句の中で，日本語 WordNet オントロジにリンク付け出来たものを各関係の評価とした．なお，本研究では，日本語 Wiktionary からデータ抽出するために，Zesch らが開発した API[3] を日本語 Wiktionary 向けに拡張し，使用している．

7.1　評価結果

語彙量についての評価結果を，表 2 に示す．まず，日本語 Wiktionary から単純に抽出した 31,302 語句 (ページ) を，日本語 WordNet オントロジが持つ 93,834 語句と比較すると，18,660 語句追加可能であることが分かった．文法オントロジを活用すると，派生語や漢字・読み表記関係にある語句が増加し，50,657 語句となり，その内 37,073 語句が追加可能となる．よって，合計の語彙量としては，130,907 語句となり，単純に抽出した場合の 112,494 語句の約 1.16 倍となった．

品詞・活用形情報や各種関係の評価結果を，表 3 に示す．リンク抽出数とは，日本語 Wiktionary から抽出して文法オントロジに登録できたリンク数である．リンク付与数とは，抽出したリンクを活用し，日本語 WordNet オントロジの語意に対して付与できたリンク数である．

品詞・活用形については，日本語 Wiktionary から 36,243 語句に対して抽出できており，日本語 WordNet オントロジの 33,909 語意にリンク付与できた．全体で，158,058 語意あるので，約 21% の語意に対して，品詞・活用形情報を登録できたことになる．

派生語関係の評価結果については，日本語 Wiktionary から 3,313 個の派生語関係のリンク数を抽出した．しかし，日本語 WordNet オントロジの語意に対しては，317 個のリンク数に留まる．

漢字・読み表記関係の評価結果については，27,756 個のリンクを抽出し，活用によって 28,032 個を付与することが出来た．日本語 WordNet において，漢字語句を持つ語意は 127,000 個あるため，約 22% について漢字・読み表記関係が把握出来ることになる．

7.2　考察

本評価結果により，入力した語句に対する語意や，派生語，別の表記へと辿りやすくなったと言える．語彙量については，一般的な辞書が約 20 万語持つとすれば，その半数を超えることが出来ている．残りの約 7 万語に対応するには，派生語関係を増やすか，他のデータとの

表 2　語彙量の評価結果

	日本語 WordNet	日本語 Wiktionary 文法オントロジ活用前	日本語 Wiktionary 文法オントロジ活用後
抽出語句数	93,834	31,302	50,657
追加可能語句数	-	18,660	37,073

表 3　リンク付与の評価結果

	リンク抽出数	リンク付与数
品詞・活用形	36,243	33,909
派生語関係	3,313	317
漢字・読み表記関係	27,756	28,032

連携によって，補完することを検討する．

品詞・活用形および派生語関係を抽出できた語句であれば，少なくとも派生語の表記を取得することが可能となった．その中で，日本語 WordNet が持つ語意とリンク付けされていれば，ある語意の派生語関係にある語意へと辿ることもできるため，その派生語の同義語まで辿ることも出来る．しかし，前述した結果の通り，語意間の派生語関係のリンクは，317 個に限られる．これは，今回の活用では，派生語の語意まで生成しておらず，日本語 WordNet にて予め定義された派生語を対象としたためである．よって，活用するためには，語句だけでなく，派生語の語意を定義する方法を検討しなくてはならない．また，今回の文法的に生成した派生語は，必ずしも登録すべきとは限らない，という問題もある．例えば，動詞 “増える” の連用形は，“増え” であるが，一般的には名詞として使用されない．

漢字・読み表記関係については，日本語 Wiktionary の約 77% の語句について抽出できており，前述の通り日本語 WordNet の約 22% の語意についてリンクを付与できた．ただし，読み方によって意味が異なる漢字について，日本語 WordNet の語意に対しては，反映できていない．

上記の通り，日本語 Wiktionary から抽出した文法オントロジを活用することによって，日本語 WordNet と日本語 Wiktionary の語意や語句について，用法・活用，派生語関係や表記関係を補完することが出来た．しかし，語意とのリンク付けは不十分な点が多い．これは，派生語や別表記の語意を定義していない点もあるが，日本語 WordNet と日本語 Wiktionary が持つ語意を独立した扱いとしていることも要因として大きい．つまり，本来であれば，同じ語意は同じリソースに統合すべきだが，現状は別のリソースとして登録してしまっている．これは，現状の抽出情報だけでは，語意の同定が困難であったためである．

8　関連研究

WordNet と Wiktionary を統合する技術として，[4][5] がある．これらは，WordNet と Wiktionary 間で登録されている語句に対し，意味や関連語関係などから類似度を算出することによって，語彙の同定，及び統合を行っている．また，日本語 Wiktionary ではなく日本語 Wikipedia を日本語 WordNet と統合する研究 [6][7] もある．しかし，いずれも各国語由来の品詞や活用形，それを使用した派生語関係，表記関係については表現していない．

派生語の生成について，[8] では，コーパスから収集した派生語用例を生成規則の形で記述し，その適用確率を学習している．適用確率を使用することにより，派生語らしく，使用頻度も高い語が受理される．ただし，派生語は，語幹を成す名詞と接尾語との連接に限られており，本稿のような活用語は対象としていない．しかしながら，コーパスや確率の使用による判定は，本稿の派生語の精度を上げる可能性があるため，今後検討していきたい．

また，オープンな辞書データとして，IPADIC[9] がある．現在，公式では開発が進められておらず，また類義語や反意語など関連語関係を持たないため，本稿では Wiktionary を採用した．

9　おわりに

本研究では，用法や活用，派生語関係，表記関係を表現した文法オントロジを作成し，日本語 WordNet オン

トロジに対して活用した．その結果，文法オントロジ活用前の語彙量 112,494 語から，活用後 130,907 語となり，約 1.16 倍となった．また，品詞・活用形情報について 33,909 個，派生語関係は 317 個，漢字・読み表記関係は 28,032 個のリンクを付与できた．

今後の課題として，現状は語意の統合が不十分であるため，語意の説明文や関連語などを使用して，WordNet と Wiktionary 間の語意を同定，あるいは生成し，統合する方法について検討する．また，派生語については，実際には登録すべきでない使用されない語も含まれるため，コーパスや他の辞書データと連携するなどして，検証する必要がある．

参考文献

[1] Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. Development of the Japanese WordNet. In *LREC*, 2008.

[2] Graham Klyne and Jeremy Carroll. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C recommendation, W3C, February 2004. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

[3] Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *LREC*, Vol. 8, pp. 1646–1652, 2008.

[4] Francis Bond and Ryan Foster. Linking and Extending an Open Multilingual Wordnet. In *ACL (1)*, pp. 1352–1362, 2013.

[5] John McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano. Integrating WordNet and Wiktionary with lemon. In *Linked Data in Linguistics*, pp. 25–34. Springer, 2012.

[6] 山田一郎, 呉鍾勳, 鳥澤健太郎, 黒田航, 風間淳一, 村田真樹. Wikipedia を利用した日本語 WordNet への用語追加の検討. 言語処理学会第 16 回年次大会発表論文集, pp. 948–951, 2010.

[7] 森田武史, 玉川奨, 山口高平. 日本語 Wikipedia オントロジーと日本語 Wordnet の統合 (学習およびその応用). 知識ベースシステム研究会, Vol. 96, pp. 9–14, 2012.

[8] 市丸夏樹, 中村貞吾, 宮本義昭ほか. シソーラスと確率文法による派生語解析. 情報処理学会論文誌, Vol. 36, No. 4, pp. 849–858, 1995.

[9] Masayuki Asahara and Yuji Matsumoto. IPADIC version 2.7. 0 User's Manual (in Japanese). NAIST. *Information Science Division*, 2003.