百億語のコーパスを用いた日本語の語彙・文法情報のプロファイリング

スルダノヴィッチ・イレーナ（国立国語研究所・リュブリャーナ大学）†
スホメル・ヴィット（マサリック大学言語処理センター）
小木曽智信（国立国語研究所）
キルガリフ・アダム（レクシカルコンピューティング・リーズ大学）

Japanese Language Lexical and Grammatical Profiling Using the Web Corpus JpTenTen

Irena Srdanović (National Institute for Japanese Language and Linguistics/University of Ljubljana),
Vit Suchomel (Natural Language Processing Centre, Masaryk University)
Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
Adam Kilgarriff (Lexical Computing Ltd./Leeds University)

１．はじめに

近年、一億語を超えた大規模な現代日本語書き言葉均衡コーパスが完成し、その大きなプロジェクトの成果として新しいアノテーションツール、電子化辞書、コーパス検索ツールなどの日本学以外の様々な分野に応用できるリソースが作成されてきた。次の段階として、コーパス量を増やす必要性が明らかになり、今までのデータでは十分把握できず、抽出できなかった言語的情報を得るために超大規模なウェブコーパス構築が始まった。こうした中、様々な言語でウェブコーパス作成の重要性が認識されてきて、多言語のための TenTen と呼ばれるウェブコーパス群の構築が行われている。本論文において、まず新たに作成された JpTenTen という日本語の 100 億語の超大規模なウェブコーパスを紹介する。このコーパスは、SpiderLing（Pomikalek and Suchomel 2012）などのツールでデータをクロールし、クリーニングを行った上で、MeCab と UniDic2 (小木曽ら 2011)で形態素解析し、短単位と長単位アノテーションを付与した。コーパスは Sketch Engine というレクシカルプロファイリングツール（Kilgarriff ら 2004）に搭載した。このツールは既に 4 億語の日本語コーパス JpWaC を基にした語彙･文法プロファイリングを可能にしているが（Srdanović ら 2008）、本研究によって新たに可能になった成果は以下の通りである。

- 超大規模なコーパスを構築し、スケッチエンジンツールに載せた。その結果、今までできなかった言葉の組み合わせなどの言語情報を取り出せるようになった。
- 長単位と短単位のアノテーションを利用したことで、以前より統一された短単位のデータと、以前には存在しなかった長単位のデータが利用可能になった。
- 品詞タグだけでなく、UniDic の活用形および活用型等の英訳アノテーションを利用し、以前にはなかった活用形に関する詳細な情報を取り出せるようになった。
- 「文法関係ファイル」のデータを更に整備し、今まで取り出せなかった語と語の組み合わせおよびその振る舞いの情報が抽出できるようになった。

以上の外に、２語以上の共起抽出などの新しく開発した機能により、以前にはできなかった情報習得および表示ができるようになってきた。

本論文では、第２章においてコーパスの構築を紹介した上で、第３章においてコーパスのアノテーションおよび短単位と長単位の語彙プロファイリングのメリットについて述べる。第４章は、新しい「文法関係ファイル」によって抽出できるようになった語彙・文法情報を紹介し、第５章では、具体的な例を取り出し、百億語の日本語のコーパスからどのような言語的情報が得られるかについて述べる。

† irena.srdanovic@ff.uni-lj.si

２．TenTen コーパス群と JpTenTen コーパス構築

近年、ウェブデータを用いたコーパス構築のメリットが認識され、それに関する研究が増加してきた。最初の日本語大規模ウェブコーパス JpWaC は、Baroni and Kilgarriff (2006)、Sharoff（2006）が提案した方法を利用し、WaC 群の一つとして開発されたものである（Srdanović ら 2008)。近年「Corpus Factory」（コーパスファクトリ）（Kilgarriff 2010）というプロジェクトの枠組みで、10^{10}（百億語）の TenTen という新しいウェブコーパス群の開発が始まり、さまざまな言語のコーパスが構築された。TenTen 群の一つとして日本語超大規模コーパス JpTenTen が 2011 年に作成された（Pomikalek and Suchomel 2012)。正式な名前は「JpTenTen11」[1]である。

JpTenTen は以下の手順で構築された。

（１） 日本語の言語モデル作成。日本語のウィキペディアからデータを利用し、モデル学習を行った。約 1000 ページの日本語ウェブページをさまざまなエンコーディングで取得した。（Kilgarriff 2010）

（２） 言語コーパス作成用の SpiderLing クローラー（Pomikalek and Suchomel 2012）によって、前述したモデルを利用し、日本語のウェブページをクロールした。

（３） JusText を利用し（Pomikalek 2011）、「文にあるテクストだけ」（text in sentences only）を収集し、それ以外のテクストではないデータおよび「ボイラープレート」（boilerplate）を削除した。

（４） 「オニオン」というツールで、段落レベルの情報で重複したデータを削除した（de-duplicate）（Pomikalek 2011)。

（５） 形態素解析ツール MeCab 0.98 および電子化辞書 UniDic 2.1.0 を利用し、全体のコーパスを処理し、アノテーションを付加した(小木曽ら 2011)。その際、UniDic の品詞・活用形・活用型のマッピングを行い、英訳のタグセットを作成した。

（６） Comainu 0.60 を利用し、UniDic の長単位の処理およびアノテーションを行った。このステップは時間を要するため現時点では作業中であり、サンプルコーパスが完成しているところである。

（７） 以前作成した日本語の「文法関係ファイル」を基にして（Srdanović ら 2008、スルダノヴィッチ・仁科 2008)、UniDic の英訳タグセットと正規表現を利用し、新しい日本語の「文法関係ファイル」を作成した。

（８） データの記号化（encoding）とワードスケッチのコンパイルは、Sketch Engine（Kilgarriff 2004）が利用している Manatee というシステムで行った。

UniDic の短単位でタグされた JpTenTen は、10,321,875,665 語のデータである。15,553,207 のウェブページ、734,758 のドメインからのものである。高頻度のドメインは 28,474 のウェブページからなっており、一つのウェブページからなるドメイン数は 224,293 である。表１は、コーパスにあるトップ頻度の５つのドメインを示す。

表１　コーパスにあるトップ頻度の５ドメインおよび
ドメインごとのウェブページ割合

ドメイン	Com	jp	net	info	Other
ページ割合	50%	32%	9%	5%	4%

[1] 「Jp」は、日本語を指す 2 文字のコードである（ISO-6390-1pcode)。数年後更新するモニターコーパスとして計画されているため、「11」は、2011 年にウェブから得られたデータのことを示す。

３．UniDic 短単位と長単位アノテーションを付加した JpTenTen

日本語は単語の分かち書きがなされず多様な表記法を持つため、日本語のコーパスにとって単語情報（形態論情報）のアノテーションは重要である。特に、単語の区切り方をどうするのか、多様な表記をどのようにまとめ上げるのか、という点は大きな問題となる。

JpWaC コーパスでは、従来 ChaSen 標準の辞書である IPADIC を利用してきたが、この辞書では、単語の区切り方の揺れや、表記のまとめ上げなどは言語の研究にとって十分であるとは言えない点があった。たとえば、区切り方の面では、「株式会社」が 1 語である一方で「有限／会社」「合資／会社」は 2 語に分割されるような揺れがあった。また表記の面では、「ネギ」「ねぎ」「葱」を見出し語として一つにまとめ上げることができなかった（読みとしてはまとめられるが、そうすると「禰宜」と区別されない）。

今回、JpTenTen では、UniDic を利用することによってこうした問題に対処した。UniDic は、BCCWJ の開発にあたって整備された形態素解析辞書で、このような問題を解決することができる。UniDic は、短単位と呼ばれる厳密な規定によって単語の区切り方が定められており、揺れが少ない斉一な単位による解析が可能になっている（小椋ら 2011）。また、語彙素・語形・書字形・発音形という見出し語の階層構造を持っており、利用者が必要に応じて、見出し語のレベルを選択して利用することができる（伝ら 2007）。たとえば、表記そのものに関心があるのであれば「書字形」を、語形の差異に関心があるのであれば「語形」を、辞書見出し（lemma）のレベルでまとめ上げたいのであれば「語彙素」を利用すればよい。

UniDic では、前述の「株式会社」は規程に従って他と同様に「株式／会社」と 2 語に分割され、「ネギ」「ねぎ」「葱」には共通して語彙素「葱」・語彙素読み「ネギ」の情報が付与される。さらに、新しい JpTenTen では、BCCWJ と同様に長単位による解析も行い、短単位と長単位の両方で利用することを可能にした。長単位とは、文節を基準とした語の単位で、まず文節を区切りとし、さらに文節のうちの付属語を切り出したサイズになる。また、短単位で分割される漢語サ変動詞や一部の複合辞は 1 長単位となる。次の例は、同じ文を短単位と長単位で分割した例である。

短単位：私/は/国立/国語/研究/所/で/日本/語/を/研究/し/て/いる/
長単位：私/は/国立国語研究所/で/日本語/を/研究し/ている/

短単位が辞書の見出しとしてあらかじめリストアップされたかなり短い単位であるのに対し、長単位は実際にコーパスに出現する形に基づいて作られる比較的長い単位である。ただし、多くの事例では短単位と長単位は一致する。長単位は、長単位解析器 Comainu により、UniDic を使って行われた形態素解析結果である短単位を組み上げる形で作成される（小澤ら 2011）。

JpTenTen に利用した UniDic の品詞、活用形、活用型は英訳した上でコーパスに載せた。品詞マッピングの例を表２に示す。

表２　UniDic の品詞マッピング

品詞	品詞(英訳)	記述
代名詞	Pron	pronoun
副詞	Adv	adverb
助動詞	Aux	auxiliary_verb
助詞-係助詞	P.bind	particle(binding)
助詞-副助詞	P.adv	particle(adverbial)

４．スケッチエンジンに載せた JpTenTen

４.１　コンコーダンス

JpTenTen コーパスをスケッチエンジンに搭載することにより、ツールのウェブページからアクセスができ、標準的なコンコーダンスとしての機能が利用できる。コンコーダンスは、語彙素、語句、単語、文字および CQL 機能（Corpus Query Language、コーパス検索言語）で正規表現とデフォルト属性を基にした共起、文法的パターンなどの項目の検索方法が指定できる。ここは、UniDic の短単位と長単位で分析されている語彙素で検索ができる。図１は、コンコーダンスにあるデフォルト属性の選択肢を示している。以前は単語（word）、語彙素（lemma）、タグ（tag）での検索だけが可能だったが、現在は活用形（infl_form）、活用型（infl_type）また語彙素読み（lemma_kana）で言語的情報の検索ができるようになった。

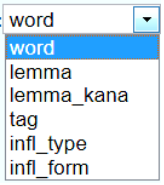

図１　コンコーダンスにあるデフォルト属性の選択肢

図２は、コンコーダンスで可能な表示の例を示している。キーワードだけのアノテーションを表示するか周りの単位のアノテーションも表示するか、またどのアノテーションタイプを表示するかを選択できる。図の例は上から順に、(1)キーワードの語彙素、(2)キーワードの語彙素と品詞、(3)キーワードの単語・語彙素・読み方・品詞・活用型・活用形、(4)キーワードとコンテクストの語彙素と品詞を表示したものである。

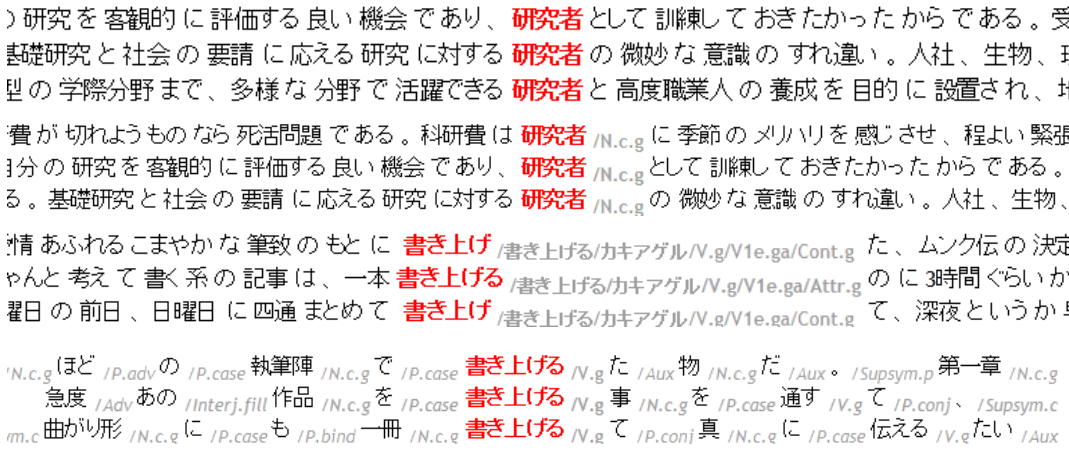

図２　コンコーダンスにある可能な表示

４.２文法関係ファイルとワードスケッチ

日本語の「文法関係ファイル」において語彙・文法的関係を決定した結果、コンコーダンスだけでなく、キーワードの語彙・文法的プロファイリング、キーワードのシソーラス、類似した語の差異と共通点などをウェブ上で１ページにまとめた言語的情報が見られる。

日本語のための「文法関係ファイル」は 2007 年に初めて作成された（Srdanović ら 2008）。ファイルの作成においては、Gahl（1998）によって提案された「corpus query syntax (コーパス検索シンタクス)」を実装し、主に品詞と正規表現を利用した。日本語の語彙・文法的規則を作成するにあたって、日本語の動詞、名詞、形容詞、副詞、接尾辞、接頭辞、助動詞

などの単位をカバーし、それぞれの品詞がどのような語と語で組み合わされ、どのようなパターンで現れるかをさまざまな言語的データから簡単に抽出・観察できるようになった。

本研究では、既存の「文法関係ファイル」を様々な面で整備・更新した。内容を以下にまとめる。[2]

(1) 第 3 章に説明した MeCab-UniDic の短単位と長単位のアノテーションを採用するため、「文法関係ファイル」に以前利用した ChaSen‐IPADIC のタグから MeCab-UniDic へのタグマッピングを行った。
(2) 品詞だけでなく、新たに活用型・活用形に基づいて正規表現で語彙・文法パターンを作成した。
(3) 以前はカバーされなかった文法関係を新しく作成した。

それぞれの改善点は、以下の「文法関係ファイル」のパターンの例、または図１と図２のワードスケッチの例に見られる。

*DUAL[3]
=modifier_Ai_cont/modifies_N+する
2:[tag="Ai.*" & word!="なく|無く" & infl_form="Cont.*"] [tag="Pref"]? 1:[tag="N.c.vs"]

語彙・文法関係は、主に 2 項（dual）関係タイプとして設定する。たとえば、以上の例は名詞‐普通名詞‐サ変可能（tag="N.c.vs"）を検索すると、それを修飾する連用形の活用形（infl_form="Cont.*"）にある形容詞が現れる（tag="Ai.*"）。また形容詞をキーワードにして検索すると、それに呼応する名詞‐普通名詞‐サ変可能の例が現れる。このパターン（いわゆる「文法関係」）に「modifier_Ai_cont/modifies_N+する」という名前を付けた。

以上に利用した省略の説明は以下のとおりである。

tag="N.c.vs"- noun.common.verb_suru の品詞省略
infl_form="Cont.*"Continuous_ren'yo の活用形省略
tag="Ai.*"　adjective –i の品詞省略

このパターンの内容には、前述した(1)のタグマッピングの結果の例、(2)の活用形の利用、(3)新しく作成した文法関係の例を含んでいる。

図３は以上のパターンを利用したワードスケッチの例を示す。たとえば、「結婚」というサ変名詞がどのような連用形の形容詞と結びつくかを検索した結果である（例えば、めでたく結婚する、早く結婚する、仕方無く結婚するなど）（図３の１欄目）。また、「素晴らしい」という形容詞が連用形の活用形の場合、どのサ変名詞と結びつくかを示した結果である（例えば、素晴らしく洗練する、素晴らしく感動する、素晴らしく調和する、素晴らしく充実するなど）（図３の２欄目）。

検索した結果は短単位であり、「する」は別の単位と扱われているため、結果にはサ変名詞だけが表示される。一方、図３の３欄目は、UniDic の長単位で検索し、「結婚する」を一つの単位として扱った例である。JpTenTen サンプルコーパスから取り出した結果なので、高頻度の組み合わせの「めでたく結婚する、早く結婚する」だけが表示されているが、長単位のキーワードで検索できるメリットがある。全体のコーパスが利用可能になると、抽出できる結果が増加する。

図３のそれぞれのの欄に表示されている数字は、1 列目がコーパスの中の共起頻度を示し、2 列目がその共起の統計的な重要度（salience）を示している。[4]

---

[2] 以前のデータの評価および問題点について Srdanović ら（2011）、Kilgarriff ら（2010）を参考されたい。

[3] 語彙・文法関係の 2 項（dual）などの設定について詳細は Srdanović（2008）らを参考されたい。

[4] 1 列目の数字をクリックすると、コーパス中にあるキーワードとそれぞれの共起語が含まれる例文がコン

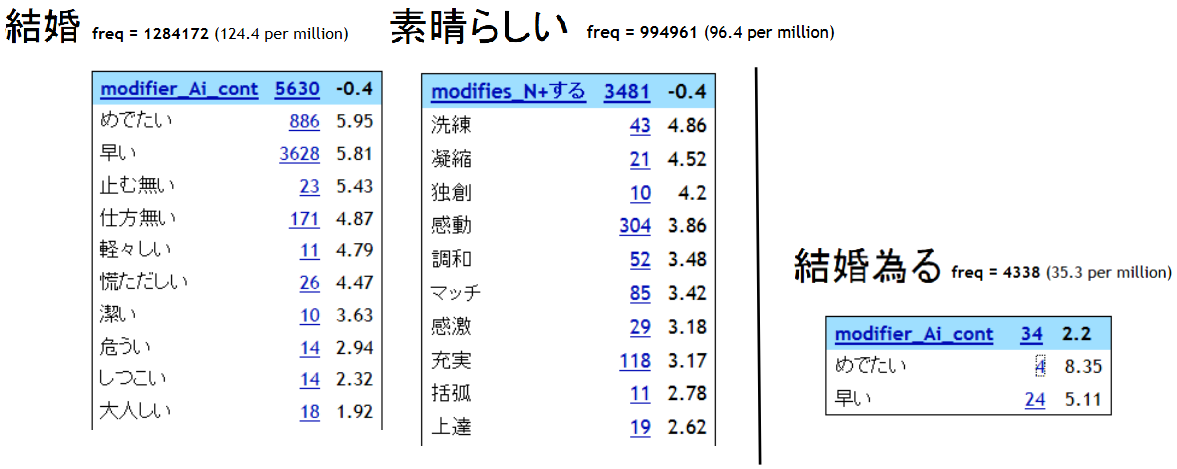

図３「結婚＋形容詞連用形」および「素晴らしい＋名詞普通名詞サ変可能」のワードスケッチの例（JpTenTen,UniDic2 短単位）。また、「結婚する＋形容詞連用形」のワードスケッチの例（JpTenTen のサンプル,UniDic2 長単位）

以上の例は、前述した長単位による情報抽出のメリットを示すもので、以前には抽出できなかった「サ変名詞＋する」の組み合わせパターン、および活用形のタグを用いた抽出の例である。

５．JpTenTen を用いた語彙・文法情報のプロファイリング

本章では、百億語の JpTenTen コーパスから取り出せる語彙・文法情報プロファイリングのいくつかの例を紹介する。

５．１　まとめた形のキーワードのプロファイリング

図４は、ワードスケッチの「女性」というキーワードの様々なパターンの例で、新しく抽出できるようになった文法関係のバラエティーを示す。パターンは、「女性＋助詞」「女性＋名詞」「女性＋の＋名詞」「名詞＋の＋女性」「女性＋に＋動詞」「女性＋が＋動詞」などである。キーワードがどのようなシンタクスの中でよく利用されているか、どの助詞と結びつくか、どの形容詞・形状詞に修飾されるかなどの細かい語彙の振る舞いが観察できる。

スペースの制限のため、それぞれのパターンの結果を省略し、一番重要度が高い 3・4 語を示した。

コーダンスの中で表示される。文法関係用語のリンク（modifies_N など）をクリックすると、その文法関係が正規表現と品詞を利用して、どのように決定されているかを確かめることができる。

スルダノヴィッチ・仁科（2008）に示したように、このような情報により、キーワードの意味を把握できるため、キーワードの意味記述などのために辞書学によく応用される。それ以外にも、言語学、言語教育などの分野に幅広く利用できる。

図４JpTenTen から取り出せる「女性」という名詞のさまざまなパターン
（パターン結果は省略した）

５．２　短単位および長単位で見る語彙プロファイリングのメリット

3 章に既に紹介したように、UniDic にはさまざまなメリットがある。本章では短単位および長単位で取り出せる言語的情報の例を上げるが、特に強調するポイントは、以下のとおりである。

- 短単位により、言語単位がどのような部分から構成されているのか、調べられること。特に派生語と関連して、接尾辞、接頭辞、非自立可能な品詞のそれぞれの特徴、振る舞い傾向を細かく調べることができる。例えば、どの形容詞が「～らしい、～こい、～臭い」などの接尾辞とよく結びつく傾向があるか、また「研究」という名詞の後ろにどの接尾辞がよく付くかといった情報が大規模なデータにより把握できる。

- 長単位で、複数の単位からできている言語単位の振る舞いを検討することができる。以前は抽出できなかったサ変動詞、複合名詞、複合動詞、のような複合語が語彙素になり、これらの語彙を単位とする組み合わせとして抽出できるようになった。これにより、長単位をキーワードとして調べることができるだけでなく、他の語をキーワードとして調べたときに長単位にもとづく情報が得られるという二つ面でメリットがある。

図５は、長単位でタグづけされたサンプルコーパスから取り出した「研究者」という名詞および「興味深い」という形容詞の例を示す。

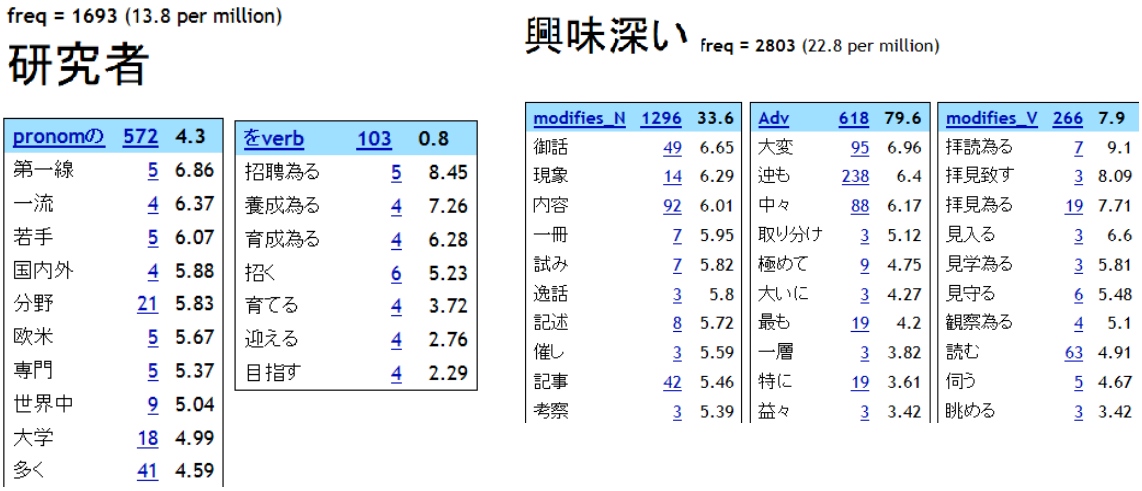

図５JpTenTen（長単位、サンプル）から取り出した「研究者」および「興味深い」のプロファイリング

これらのキーワードは短単位では取り出せなかった語であり、このような複合語のプロファイリングができるのは非常に重要である。取り出した結果にも複合語のデータが多い。例えば、「第一線の研究者、国内外研究者、世界中の研究者、研究者を招聘する」また「興味深いお話、興味深く拝見致す、興味深く拝読する」などである。

５．３　複数単位の抽出

新しく追加された機能で、それぞれのパターンにある単位からマルチワードスケッチページ（Multiword sketches）に飛ぶことができるようになった。図６はこのようなページ結果の例を示す。例えば、「最近の研究」から「最近の研究成果」、「新たな研究」から「新たな研究領域」、「とても興味深い」から「とても興味深く読む」などの複合語が並んだ例が見られる。

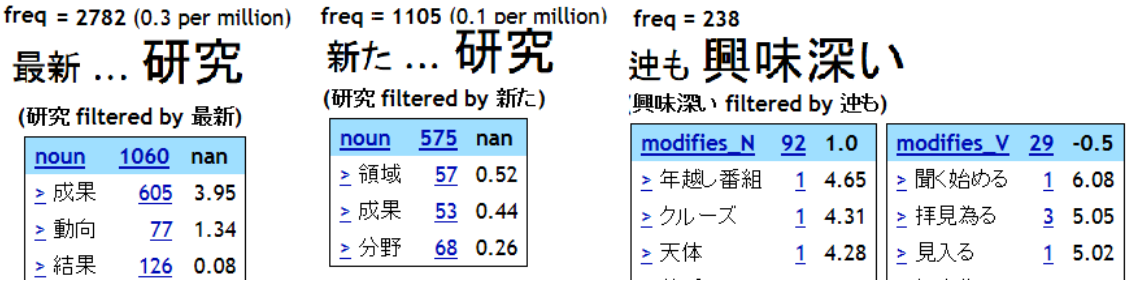

図６マルチワードスケッチの例（最近の研究～、新たな研究～、とても興味深い～）

５．４　語彙・品詞・活用形・活用型・パターンの頻度リスト

スケッチエンジンツールでは、さまざまな語彙・品詞・活用形・活用型の頻度リストが取り出せる。図７にその例を示す。1 欄目は、UniDic 短単位で解析された 100 億語のコーパスに現れる品詞の高頻度順リストである（名詞-普通名詞-一般、助詞 - 格助詞、助動詞、名詞-普通名詞-サ変可能など）。2 欄目は、もっとも高頻度の活用形（連用形-一般、終止形-一般、連体形-一般、連用形-促音便など）、3 欄目は、もっとも高頻度の活用型のリストである（助動詞-ダ、五段-ラ行、助動詞-タ、サ行変格など）。

図７の 4 欄目は、コーパスに現れる「助詞 - 接続助詞の「て」＋動詞 - 非自立可能」というパターンの頻度リストである[5]。　頻度の高いほうから、「ている、て来る、てしまう、て行く、て見る、てくれる」の順番で日本語の（テ形に接続する）補助動詞が現れる。Martin（2004、512 ページ）は、1964 年の国立国語研究所の「現代雑誌九十種の用語用字」のデータを基にして、日本語の主な補助動詞を相対頻度で並べて表に示している。現れている高頻度の補助動詞は、ほとんど図７の 4 欄目と統一している。並んだ順番もほとんど類似しているが、微妙な違いが見られる。例えば、「てしまう」は Martin（2004）の表では「行く、くれる、くださる」よりやや低い頻度で、JpTenTen のデータでは「しまう」のほうがやや頻度が高くなっている。

特定の単語、単語のグループ、一つの品詞の語彙、一つパターンなどを対象にして、超大規模コーパスから活用形、頻度などのデータを取り出すことで、今後の教育シラバス作成などに応用できる豊富な情報が得られるといえる。

tag	Freq
N.c.g	1588826682
P.case	1359247326
Aux	863345234
N.c.vs	554425638
V.bnd	532961623
V.g	506815227
Supsym.p	447290326
N.num	414314084
Supsym.c	376814114
P.conj	372440519
P.bind	355183028
Supsym.g	280555426
N.c.adv	215750266
Suff.n.g	206247617
Sym.ch	204585993
Adv	173481056
Supsym.bo	171503010
P.adv	158837049
Supsym.bc	153531627
N.c.count	128986663
P.fin	119879442

infl_form	Freq
Cont.g	607632121
Concl.g	485365140
Attr.g	469486993
Cont.t	158394931
Irr.g	116420893
Cont.ni	48221528
Cond.g	36496602
Cont.i	31451799
Vol_tent	29269791
Irr.sa	22309244
Cont.n	14146926
Concl.n	13031503
stem.g	12867278
Imp	12830635
Cont.int	6790980
Attr.n	1550658
Irr.se	1347788
Cond.int	1187705
Real.g	683018
Attr.abbr	642454
Cont.u	640659

infl_type	Freq
Aux.da	260552830
V5.ra	233862083
Aux.ta	214041705
sa_irr	196667604
Adj	167134814
V5.wa_a	131979751
V1i.a	110162200
Aux.masu	96666486
Aux.desu	83078022
V5.ka	73080756
Aux.nai	54693529
Aux.reru	52457777
V5.sa	44028481
V1e.ta	37135919
V1e.ra	35918383
V1e.ka	29983141
V5.ma	26376522
Aux.nu	23359675
V1i.ka	22916985
V1e.ma	22535426
V1i.ma	21574027

lemma	Freq
て 居る	5259778
て 来る	1107200
て 仕舞う	716055
て 行く	551444
て 見る	509031
て 呉れる	506095
て 下さる	331408
て 貰う	251620
て 頂く	193982
て 置く	183006
て 有る	115184
て 遣る	98028
て 上げる	63500
て 参る	16593
て 見せる	11584
て 為る	10239
て いらっしゃる	9846
て 回る	8578
て 出来る	7959
て 成る	7727

図７JpTenTen における品詞・活用形・活用型・パターンの頻度リスト（UniDic 短単位）

６．まとめ

本論文では、新規の超大規模な日本語のウェブコーパス JpTenTen の構築とそのアノテーションを紹介した上で、百億語のコーパスを用いた日本語の語彙・文法情報のプロファイリングの実例を紹介した。日本語の様々な語彙・文法情報を UniDic の短単位と長単位の品詞・活用形・活用型に適用することで、以前できなかった語と語の振る舞いの情報を抽出できるようになってきた。この成果は日本語学、対照言語学、日本語辞書学、日本人学習者用英語辞書学、日本語教育、日本語言語処理、心理学などの研究分野に活用できると期待される。

[5] 利用した検索パターンは[tag="P.conj"& word="て"] [tag="V.bnd"]

謝　辞

本研究は、博報財団第 7 回「日本語海外研究者招聘事業」による研究「日本語教育における語の共起関係」（平成 24～25 年度、受入機関：国立国語研究所、招聘研究員：スルダノヴィッチ・イレーナ）およびチェコ教育科学所によるプロジェクト「LINDAT-Clarin LM2010013」（研究員：スコメル・ヴィット）の補助を得ています。

文　献

スルダノヴィッチ・イレーナ, 仁科喜久子（2008）「コーパス検索ツール Sketch Engine の日本語版とその利用方法」『日本語科学』23 号, 国書刊行会, pp. 59–80.

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（2007）「コーパス日本語学のための言語資源 : 形態素解析用電子化辞書の開発とその応用」『日本語科学』22, pp. 101-123.

小椋秀樹・小磯花絵・冨士池優美・宮内左夜香・小西光・原裕（2011）国立国語研究所内部報告書『『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（上・下）』

小澤俊介, 内元清貴, 伝康晴（2011）「BCCWJ に基づく中・長単位解析ツール」, 特定領域「日本語コーパス」平成 22 年度公開ワークショップ予稿集, pp. 331-338.

小木曽智信・伝康晴（2011）「UniDic2.0:言語資源としての電子化辞書」特定領域研究「日本語コーパス」平成 22 年度全体会議予稿集, pp. 411-418.

Baroni, Marko and Kilgarriff, Adam (2006) Large linguistically-processed Web corpora for multiple languages, In Proceedings EACL Trento, Italy

Gahl, Susanne (1998) Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus, ms., ICSI-Berkeley

Kilgarriff, Adam, Rychly, Pavel, Smrž, Pavel and Tugwell, David (2004) The Sketch Engine. Proceedings of EURALEX. France: Université de Bretagne. pp. 105-116.

Kilgarriff, Adam, Kovář, Vojtěch., Krek, Simon, Srdanović, Irena, Tiberius, Carole (2010) A Quantitative Evaluation of Word Sketches. Proceedings of the XIV Euralex International Congress. Leeuwarden:Fryske Academy. pp. 7

Kilgarriff, Adam, Reddy, Siva, Pomikálek, Jan and Pvs, Avinesh (2010) A corpus factory for many languages. In proceedings of LREC, Malta

Martin, Samuel E. (2004) A reference grammar of Japanese. University of Hawai'i Press, Honolulu

Pomikálek, Jan (2011) Removing Boilerplate and Duplicate Content from Web Corpora. PhD thesis, Masaryk University, Brno

Pomikálek, Jan, Suchomel, Vít (2012) Efficient Web Crawling for Large Text Corpora. ACL SIGWAC Web as Corpus (at conference WWW)

Sharoff, Serge (2006) Open-source corpora: using the net to fish for linguistic data, International Journal of Corpus Linguistics, 11 (4), pp. 435–462.

Srdanović, Irena, Erjavec Tomaž and Kilgarriff, Adam (2008) A web corpus and word-sketches for Japanese. Shizen gengo shori (Journal of Natural Language Processing) 15/2. pp. 137-159.

Srdanović, Irena, Ida, Naomi, Shigemori Bučar, Chikako, Kilgarriff, Adam, Kovář, Vojtěch (2011) Japanese Word Sketches: Advantages and Problems. Acta Linguistica Asiatica, 1 (2), pp.63-82.

関連 URL

国立国語研究所の言語コーパス整備計画 KOTONOHA　http://www.ninjal.ac.jp/kotonoha/

スケッチエンジンツール Sketch Engine　http://www.sketchengine.co.uk/

クローラ SpiderLing　http://nlp.fi.muni.cz/trac/spiderling

Comainu に関する参考文献 https://maro.ninjal.ac.jp/Comainu/related_paper/

形態素解析辞書 UniDic http://download.unidic.org/

MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://mecab.googlecode.com