

# コーパス検索ツール Sketch Engine の日本語版とその利用方法

SRDANOVIĆ ERJAVEC Irena, 仁科喜久子

(東京工業大学)

キーワード

Sketch Engine, コーパス言語学, 辞書学, 第二言語教育, 共起

## 概要

近年コーパス構築と利用に関してのさまざまな研究が展開しているが、本稿ではコーパス検索ツール Sketch Engine の日本語版作成と利用方法について報告する。標準的なコーパス検索ツールと異なる点は、コンコーダンス機能以外に語に付随する文法とコロケーション情報を Web 上の 1 頁にまとめる“Word Sketch”機能を持ち、シソーラス情報や意味的に類似する語の共通点と差異を示す“Thesaurus”と“Sketch Difference”機能も含むことである。現在の Sketch Engine 日本語版は JpWaC という 4 億語の大規模 Web コーパスを有しており、他のコーパスを搭載することも可能である。本稿では、Sketch Engine によるコーパス利用の例として日本語学習辞書に焦点を当て、さらに日本語学研究、日本語教育などへの応用の可能性について述べる。

## 1. はじめに

コーパスが言語研究の分野に普及し始めた 1980 年頃には、コーパスを利用すべきかどうかということが議論されていた。ほぼ 10 年後には、コーパスの規模や代表性の問題へとその議論が移った。大規模なデータの取り扱いが以前ほど難しくなくなっている現在、大規模コーパスから適切なデータをできるだけ早くどのように抽出するかの議論が行われている。80 年代にコーパスからデータを抽出するためのコンコーダンス・ツールが開発されているが、そのツールは既に「伝統的なツール」だと考えられている (Kilgarriff & Rundell 2002)。大規模なコーパスを検索対象とすると、一つのキーワードに対して 500 例、1,000 例、20,000 例以上と用例が大量に表示されるため、その大量な例を簡単に扱うことがむずかしい。そこで 2000 年頃、コンコーダンス・ツール以外の機能を含んだコーパス検索ツールの開発が始まった (Heid et al. 2000, Kilgarriff & Tugwell 2001)。本報告では、その中の一つである Sketch Engine (Kilgarriff et al. 2004) を取り上げ、Srdanović et al. (2008 予) が作成した日本語版とその利用の可能性を紹介する。

Sketch Engine はコンコーダンス機能以外に語に付随する文法とコロケーション情報の記述を Web 1 頁にまとめる機能を持ち (Word Sketch)、さらにシソーラス情報 (Thesaurus) や類義語間の共通点と差異の提示 (Sketch Difference) など、言語的な情報をコーパスから取得するための新しい方法を提供する。英語をはじめとする他の言語のバージョンは既にコーパス言語学・コーパス辞書学・第二言語教育の分野においてここ数年利用されている。日本語バージョンでも日

本語の研究や資源・教材作成に利用できる環境が整っている。そこで本稿では、Sketch Engine の日本語バージョンの有効な利用方法について紹介する。

## 2. Sketch Engine の日本語化とその機能

コーパス検索ツール Sketch Engine は最初に英語のために作成され (Kilgarriff et al. 2004), その後、他の言語バージョンが追加されたものである。日本語版は Erjavec et al. (2007) によって作成された 4 億語 (厳密に言えば延べ形態素数 4 億) という大規模 Web コーパスを搭載しており、今後もその他のコーパスを搭載することが可能になっている。

以下では、日本語版を実装した時の主なステップを紹介し、Web コーパスに関する議論や最近の研究結果について述べる。ここではコーパスを Sketch Engine に搭載する過程、「文法関係」ファイルの作成方法について簡単に述べた上で、Sketch Engine の主な機能を紹介する。

### 2.1. Sketch Engine へのコーパス搭載

Web からログインできる Sketch Engine のツール (<http://www.sketchengine.com>) に、まず 4 億語を有する JpWaC という Web コーパスを搭載した (図 1)。このコーパスは、Sharoff (2006), Ueyama & Baroni (2005) などの方法により、Web から約 5 万ページを収集し、WAC (Baroni & Bernardini, eds. 2006) と BootCat ツール (Baroni et al. 2006) によってデータを整理し、作成された。できる限り文章のみのデータを対象とし、HTML などの不必要な文字列を除去する「ボイラープレート削除 (boilerplate removal)」を実施した。これにより、Web ページから定型的なタグ、ナビゲーションフレーム、コード、リンクなどのテキストではないデータを削除することができた。不必要な文字列が削除されたデータは形態素解析ツール ChaSen によって、token (出現した形のまの単語・形態素)、lemma (活用形を含む単語・形態素の代表)、tag (Erjavec et al. 2006 による英訳した日本語品詞) などに分析済のものである。収集されたデータは二つのドメイン (.jp と .com) に基づいており、その中で最も多いのはブログからのデータである (Erjavec et al. 2007, Srdanović et al. 2008 予)。

Sketch Engine にコーパスを搭載することにより、標準的なコンコーダンスとしての機能が利用できる。図 2 は語句、共起、文法的パターンなどの検索画面を示している。図 3 は広範囲な基準によるコンコーダンスのソート、ランダムなコーパスサンプルなどの機能を示すコンコーダンスの画面である。コーパスデータをジャンル別に分類することは今後の課題と考えられるが、現時点では、ユーザーは検索結果を見ながら、オリジナル URL のページ・リンクに接続でき、ジャンルなどの情報に関して、自分で判断できるようになっている。用例の引用などに関する著作権は一般の Web の利用と同じように考えられる。

今後はさらに JpWaC 以外の他の日本語コーパスが搭載される可能性もあり、「現代日本語書き言葉均衡コーパス」のような均衡コーパスの実装も考えられる (投野 2007)。

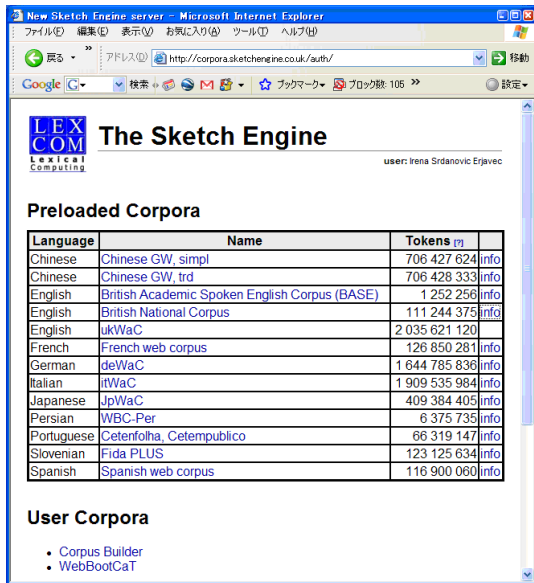


図 1 Sketch Engine の多言語版

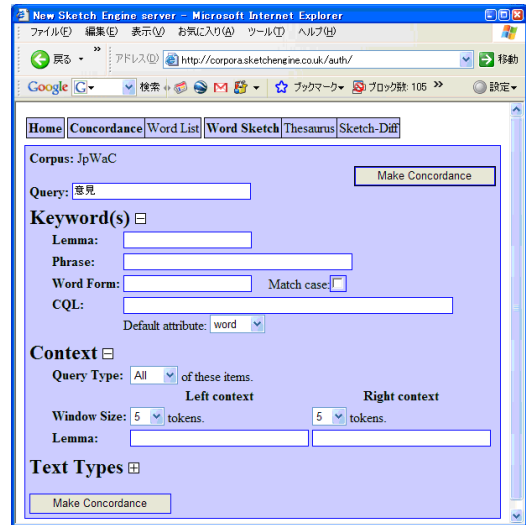


図 2 Sketch Engine のコンコーダンス



図 3 Sketch Engine のコンコーダンスによる検索結果

## 2.2. 日本語の「文法関係」ファイルと Word Sketches

日本語化の次のステップとして、文法的な関係を決定し、そのセットを載せることがある。

このために日本語文法規則が 22 項目登録され、その規則に基づいてキーワードと他の単語の可能な関係を判別することで、さらに発展的なコーパス検索方法を提供することが可能になっている (Word Sketch, Thesaurus と Sketch Difference)。

文法的な関係は Chasen の品詞に基づいて正規表現で作成され、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)」を実装している。図 4 の左側の例は「幸せな」という形容動詞をキーワードとして検索した Word Sketch の結果の一部を示している。「幸せな」という語がどのような名詞を修飾するかを検索した結果である (例えば、幸せな気分、幸

せな結婚, 幸せな家庭など)。図4の右の例は, どのような形容動詞が「家庭」という名詞を修飾するかという検索の結果を示している (例えば, 裕福な家庭, 円満な家庭など)。

図4の左右それぞれの欄に表示されている2列の数字は, 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。表中の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の含まれる例文がコン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が確かめられる。

modifies_N	3795	13.9	modifier_Ana	812	4.9
気分	448	9.01	裕福	154	11.02
結婚	156	8.14	円満	10	7.89
金持ち	76	7.97	平凡	17	7.74
ひととき	29	7.5	幸せ	156	7.74
家庭	156	7.22	貧乏	16	7.47
人生	217	7.12	温か	6	7.44
日々	73	6.87	平穏	11	7.39
ひと時	15	6.75	暖か	6	7.07
気持ち	241	6.72	幸福	38	6.97
生き方	40	6.56	健全	25	6.57
成功	61	6.48	不幸	21	6.25
貧乏	16	6.47	穏やか	9	5.99
住まい	26	6.46	快適	12	5.95
ろう	15	6.4	熱心	10	5.89
サラリーマン	20	6.25	複雑	34	5.86
結末	16	6.22	豊か	29	5.52
新婚	9	6.15	平和	18	4.58
瞬間	28	6.0	立派	7	4.52
老後	11	5.98	優秀	5	4.09
毎日	30	5.92	さまざま	12	3.75
夫婦	27	5.89	特殊	5	3.73

図4 「幸せ(な)+名詞」, 「形容動詞+家庭」の検索結果

図4で示した文法関係は2項 (dual) 関係として設定されている。形容動詞を検索するとそれが修飾する名詞が現れ, また名詞を検索するとそれに呼応する形容動詞が現れるように設定されている。この場合の文法関係は, 以下のような形で記述されている。

\*DUAL

=modifier\_Ana/modifies\_N

2:"N.Ana" "Aux" "Pref.\*"? 1:[tag="N.\*" & tag!="N.Suff.\*" & tag!="N.bnd.\*"]

上式の modifier\_Ana は形容動詞が修飾語, modifies\_N は名詞が被修飾語になる関係を示している。式の下の方の 2:"N.Ana" "Aux" "Pref.\*"? は形容動詞語幹 (N.Ana) の後に助動詞 (Aux) が来て, その後に接頭詞 (Pref.\*) が来る可能性があることを表している。1:[tag="N.\*" & tag!="N.Suff.\*" & tag!="N.bnd.\*"] は名詞の抽出法を示している。名詞 (N.\*) は, 名詞のタグが多すぎるのでその中から名詞非自立 (N.Suff.\*) と名詞接尾 (N.bnd.\*) を除く。この定義から「幸せ-な-ご-気分」, 「幸せ-な-気分」のように「ご気分」も「気分」も抽出することができる。使用した正規表現の量子子は下記の通りである。

- \* 直前の表現が 0 個以上あることを示す（以上の例では特定のタグのサブタグを含むために用いる。例えば、N.\*は N.g, N.Prop などの品詞タグを含む。）
- ? 直前の表現が 0 個か 1 個あることを示す
- ! 指定の表現を除く
- & 直前と直後の表現の接続を示す

ユーザーが正規表現を理解していなくても、Sketch Engine の画面から効率的に検索を実行することは十分可能である。しかし、正規表現を利用して Concordance の CQL 機能 (Corpus Query Language, コーパス検索言語) だけで検索できる情報もある。

以下では、いくつかの検索方法について、具体例を示す。

- 一つの語についての複数の書き方を含む検索：[word="きれい"| word="綺麗"],  
ChaSen によって 2 項目以上に分析された単語・慣用句・サ変動詞・文法パターンなどを検索：  
[word="気"] [word="に"] [lemma="する"] (3.2 章参考)

- 品詞によって検索する：[tag="N.\*"]&[ word ="つる"]

以上の項目は Word Sketch 機能でも検索できるようにすることが次の改訂版の課題である。

また、Sketch Engine は ChaSen (利用している辞書は IPADIC) の形態素解析結果を利用しているため、検索するに当たっては、IPADIC の品詞体系を確認した上で Sketch Engine を検索する方がよい。Web で公開されているツール「茶漉」でも利用できる

(<http://tell.fll.purdue.edu/chakoshipub/index2.html>)。例えば、形容動詞「幸せな」は ChaSen で「幸せ」+「な」に分けているので、「幸せ」という形式で検索すべきである。図 5 は ChaSen 解析結果を表している。品詞の英語訳は ChaSen には示されていないが、日本語版の Sketch Engine で使用している英訳品詞名を図の最右欄に付け加え示した。

token	kana	lemma	POS tag (品詞)	POS tag-eng (英語の品詞)
とても	トテモ	とても	副詞・助詞類接続	Adv.P
幸せ	シアワセ	幸せ	名詞・形容動詞語幹	N.Ana
な	ナ	だ	助動詞 特殊・ダ 体言接続	Aux
気分	キブン	気分	名詞一般	N.g
でし	デシ	です	助動詞 特殊・デス 連用形	Aux
た	タ	た	助動詞 特殊・タ 基本形	Aux
。	。	。	記号・句点	Sym.p

図 5 ChaSen による分析例「とても幸せな気分でした。」

ここで ChaSen における IPADIC 長所・短所について述べておく。ChaSen の解析単位は非常に細かいので、文法関係を定義する際には役立つ。所望のパターンを定義するためには精密なタグを用いたほうが楽で、制約条件を書く必要が少なくなるからである。一方で、検索したい表現が ChaSen によって 2 単位以上に分析された場合には、複雑なパターンで検索することになり、

Word Sketch で抽出しにくくなる。例えば、「女の子」は ChaSen によって 1 単語となっているので、Word Sketch でも簡単に検索できる。一方、「女の人」は三つの形態素に分けられるため、Word Sketch では検索できず、Concordance の中でパターンとして検索しなければならない。現在の形態素解析ツールの精度は 100 パーセントではなく、Word Sketch の中にも誤解析がまれにみられる。例えば、ChaSen が「温泉へ行った (いった)」を「温泉へ行った (おこなった)」と誤解析する例が見られた。また、Web データには方言、話し言葉もあり、これらの言語情報は未知語として解析されてしまう。これらは解析ツールの今後の課題となっている。

### 2.3. Thesaurus と Sketch Difference 機能

Thesaurus・Sketch Difference の機能は「shared triples (共有 3 元)」に基づいて、統語的に似た振舞いをする単語を自動抽出する技術を用いている。例えば、「雑誌」と「本」は「?を読む」という述語と共起し、triple 構造を成している。この技術によって類義語・対義語などを提示することができる。その基本概念については Srdanović et al. (2008 予) に述べているが、ここではその検索結果を示すために例を挙げる。

まず、Thesaurus 機能で「幸せ」という例を探すと、意味的に類似している語 (幸福, 楽しい, 喜ぶなど), および反対の意味を持っている語 (危険, 不安, 孤独など) が図 6 のように類似度を表す数値とともに表示される。

また、Sketch Difference 機能では語と語の振舞いの共通点と差異が調べられる。例えば図 7 は「女の子」と共起する語、図 8 は「男の子」と共起する語を表している。このそれぞれの表から「女の子」は「きれいな」、「かわいい」、「美しい」と共起し、「男の子」は「ハンサム」、「カッコいい」と共起する傾向があり、表現の差があることがわかる。高頻度共起語として顕著で興味深いのは「強い女の子」と「弱い男の子」である。これは、一般的に女の子は弱い・男の子は強いという社会通念・固定観念のイメージの反対であるからこそ、言及する内容として価値があるためだと考えられる。さらに、「女の子」の頻度 (16,309) が男の子の頻度 (6,486) より 2.5 倍多い。これは「女の子」が語として使用できる対象者の年齢幅などが広いためだと考えられる。また、Web 上では男の子より女の子は話題になることが多いとも考えられる。

## 幸せ JpWaC freq = 24093

幸福	0.37	不幸	0.247
健康	0.259	平和	0.258
元気	0.245	自由	0.218
不安	0.224	安全	0.183
夢	0.186	自然	0.182
希望	0.178	疑問	0.164
気持ち	0.16	興味	0.151
思い	0.147	思い	0.147
大切	0.214	不思議	0.203
いい	0.194	いい	0.175
大事	0.175	大変	0.174
若い	0.165	よい	0.165
好き	0.164	悪い	0.164
素晴らしい	0.156	すごい	0.155
大好き	0.153	面白い	0.148
新しい	0.147		
豊か	0.208	素敵	0.197
優しい	0.17	明るい	0.152
静か	0.151		
喜び	0.201	感動	0.171
楽しみ	0.163	満足	0.149
人生	0.185	生き方	0.162
生活	0.156	暮らし	0.153
成功	0.183	価値	0.164
成長	0.16		
嬉しい	0.18	悲しい	0.179
嫌い	0.179	嫌	0.173
忙しい	0.161	辛い	0.152
寂しい	0.15	怖い	0.147
愛	0.177	命	0.158
生命	0.153		
孤独	0.154		
確か	0.15		
危険	0.149		

図 6 Thesaurus での検索例：「“幸せ”」

"女の子" only patterns											
modifier_Ai	1021	10.0	がverb	2255	4.9	modifier_Ana	401	4.4	pronom	3903	3.5
思しい	5	7.0	差し出す	5	4.8	キュート	12	8.4	の		
かわいらしい	7	6.7	しゃべる	10	4.6	不細工	6	8.1	赤毛	14	6.7
ちっちゃい	5	6.7	脱ぐ	5	4.4	小柄	9	8.0	後輩	24	6.7
美しい	14	3.8	話しかけ	6	4.3	地味	7	6.2	金髪	11	6.3
明るい	5	3.4	踊る	8	4.0	綺麗	13	5.9	ふつう	13	6.1
すごい	8	2.9	映る	6	4.0	きれいな	20	5.7	同年代	9	6.0
強い	14	2.5	遊ぶ	16	3.9	上手	5	4.9	ティーン	8	5.9
良い	17	2.0	いたる	5	3.6	活発	6	4.8	同僚	19	5.9
悪い	9	1.8	好む	5	3.6	残念	5	3.7	店員	18	5.9
大きい	9	1.7	歌う	12	3.5				クラスメート	8	5.8
新しい	9	1.6	叫ぶ	6	3.5				人組	8	5.7
高い	8	0.9	亡くなる	6	3.4				レジ	11	5.7
									仲良し	7	5.6

図 7 Sketch Difference での検索例：「“女の子” only pattern」

"男の子" only patterns											
modifier_Ai	297	7.6	はverb	399	3.8	がAdj	128	2.6	prefix	17	0.2
カッコイイ	8	7.5	似る	5	2.0	ほしい	5	4.0	元	11	2.8
かっこいい	10	6.8				のpronom	756	1.8			
弱い	5	3.2	自閉症	6	6.1	節句	10	8.0			
がverb	961	5.5	白人	5	5.3	母	12	4.1			
寄る	21	5.6	番目	5	4.1	脳	6	3.5			
頑張る	5	1.9	半	7	3.3	にverb	302	1.3			
与える	5	0.2	はAdj	108	3.2	恵まれる	6	4.2			
modifier_Ana	188	5.4	高い	5	0.3	をverb	523	1.2			
ハンサム	21	10.0				殺す	6	2.3			
やんちゃ	10	9.2									

図 8 Sketch Difference での検索例：「“男の子” only pattern」

## 2.4. コーパス種類としての Web データ

英語をはじめとする多くの言語において、Web 上のデータは近年様々な技術や方法により大規模な言語資源として利用されるようになってきている。最近の研究では Web は言語学的に有益な情報

を提供できると考えられている。

一方、言語学的な分析対象としての Web データに関して様々な批判も聞こえる。その一つはノイズが多いという批判である。しかしながら、Web データについての欧米における研究では、データの量が大规模になるほどノイズの割合は問題にならないほど小さくなり、その結果は人間の主観的判断と合致することが明らかになった (Keller & Lapata 2003)。もう一つの批判は、Web のデータが片寄っているというものである。現在の Web コーパスでは、JpWaC も含めて、そのデータ内容はジャンル別に分類されていない。しかし、Web コーパスに関する最近盛んになっている研究では Web データの分類方法が検討されており (Sharoff 2006, Ueyama & Baroni 2005)、データを分類・整理することで目的・用途による分析が可能になっている。さらに、多数の言語において Web コーパス・新聞コーパス・均衡コーパスをそれぞれ比較すると、Web コーパスのほうが新聞データより均衡コーパスの結果に近づいている。Web コーパスは一般的な言語の様相を反映し、しかも量が多いので、幅広く、より良い結果が得られる (Sharoff 2006, Ueyama & Baroni 2005)。

英語の均衡コーパスと Web コーパスの比較について次のような興味深い結果が見られる。均衡コーパスは第三人称代名詞、過去形、語りのスタイル (narrative style) が多く、一方、Web コーパスは第一・第二人称代名詞、現在形・未来形、対話式 (interactive style) が多い。一般的に信頼性は均衡コーパスの方にあるといわれるが、どのような種類をどの程度含めば適切な均衡コーパスとなるのかという議論はいまだ決着を見ていない。ある言語において均衡コーパスが存在する場合には、Web コーパスと比較することによって、それぞれのコーパスの特徴が示され、均衡コーパスの構築を支援することができるという点において、Web コーパスは貴重な指針ともなる。一方、様々な理由で均衡コーパスを作れない言語がある場合、Web コーパスを代替として利用することは非常に有効だと考えられる (Ghani et al. 2001)。

もちろん、研究の目的によって、高頻度データだけに注目するのではなく、かつそれぞれの例もチェック・整理する必要がある場合には、人間の判断も必要となる。他のコーパスでも同様であるが、Web コーパスは言語規範が緩いので、他のコーパスに比較して、言語の変化が大きく、多様な様相を呈していると考えられる。規範性の面から見ると、Web データを使うのは良くないという判断もあり得るが、実際は Web データの中にも言語規範に沿ったデータも多く見られる。特に高頻度で抽出されたデータは規範性の高いものが多いと推測される。

高頻度以外の正しい言語標本が必要なら人間の判断でそれを取り出す方策もある。また、Web データは実際の言語使用と言語の生成に関わる側面を示しており、貴重なデータと考えられる。言語学者の間でも新しいメディアタイプとして Web の言語学的な役割を否定してはいけないという指摘もある (Crystal 2006)。

上記のことから、Web データの主な役割として、次の二点にまとめることができる。

- Web データは自然言語を研究するための言語学的な資源である。
- Web データは新しいメディアタイプとしてその性質を検討するための資源である。



以上、英語をはじめとする多言語の Web データについて述べた。日本語についても、青空文庫、新聞コーパス、政府白書コーパス、話し言葉コーパス、教科書コーパスなど使用可能なコーパスがあり、相互の比較をする必要があるが、これについては別稿に譲る。

### 3. Sketch Engine 日本語版の利用

本章では、まず辞書学の分野に焦点を当て、日本語版の Sketch Engine のデータが様々な辞書の作成にどのように利用できるかを述べ、次に他の言語研究、第二言語学習などへの応用を簡単に紹介する。

#### 3.1. 辞書学と Sketch Engine

電子コーパスが出現した後、辞書編纂方法には各年代において次のような開発の特徴がみられた。

- 1) 80年代には『コウビルド (Cobuild) 英英辞典』を最初として、電子コーパスやコンコーダンスが辞書編纂に用いられるようになった。
- 2) 90年代にはコロケーション統計情報として、Church & Hanks (1989) が相互情報量(MI)を提案し、辞書編纂に利用され始めた。
- 3) 2000年代には、Word Sketch のような単語の振舞いの概略を提供するツールが現れ、辞書の編纂に用いられるようになった。

以上の展開は英語の辞書学に限られているが、日本語の辞書編纂にもある程度コーパスやコーパスツールが利用され始め、将来に向けてその方法の開発が計画されている。

Sketch Engine は始めに、英語の BNC (British National Corpus) を用いて、『マクミラン英語学習辞典』の編纂に利用された (Rundell, ed. 2002)。このプロジェクトの経緯は Kilgarriff & Rundell (2002) に詳しく説明されており、コーパス辞書学の方法としては従来のコンコーダンスと比較して Word Sketch のほうが有益だと述べている。Word Sketch プロジェクトの最初の目標は、確固としたコンコーダンスラインスキャンの手法を追加し、質の高い一般性のある共起についての情報をできるだけ体系的に供給することであった。また、編集時間を短縮し、かつ良質の辞書を編纂するというのも大きな目的の一つであった。マクミラン辞書編纂プロジェクトにより、コンコーダンスが辞書編纂に重要な役割を持つと再確認できたが、一方で、辞書学者にとっては、Word Sketch が共起の抽出のみならず、意味分析を優先する出発点の役割を持つようになった。Sketch Engine の様々な機能は文法的な関係に応じて共起リストをまとめるとともに、言葉の振舞いの主要な特徴を示し、言葉の意味の把握に貢献できる。辞書編纂過程での Word Sketch の利用は、辞書学におけるコーパスデータの利用方法に大きな変化をもたらし、今後のコーパス辞書学の進展開発に影響を与えられられる。

英語バージョンが辞書編纂に貢献したように、Sketch Engine の日本語版は今後日本語の様々な種類の辞書編纂に応用できると考えられる。本節では、語彙意味論的ないくつかの分析例を取

り上げつつ、検討を進める。まず、Kilgarriff & Rundell (2002) が分析例としてあげた英語の ‘challenge’ に対応する日本語の「挑戦」の例を見る。次に、日本語学習者のための最初のコーンケーション辞典である『日本語表現活用辞典』(姫野 2004) からいくつかの例を取りだして、Sketch Engine の結果と比べる。

### 3.1.1 「挑戦」の例

図 9 は「挑戦」という言葉を入力したときの Word Sketch の結果を示す画面である。

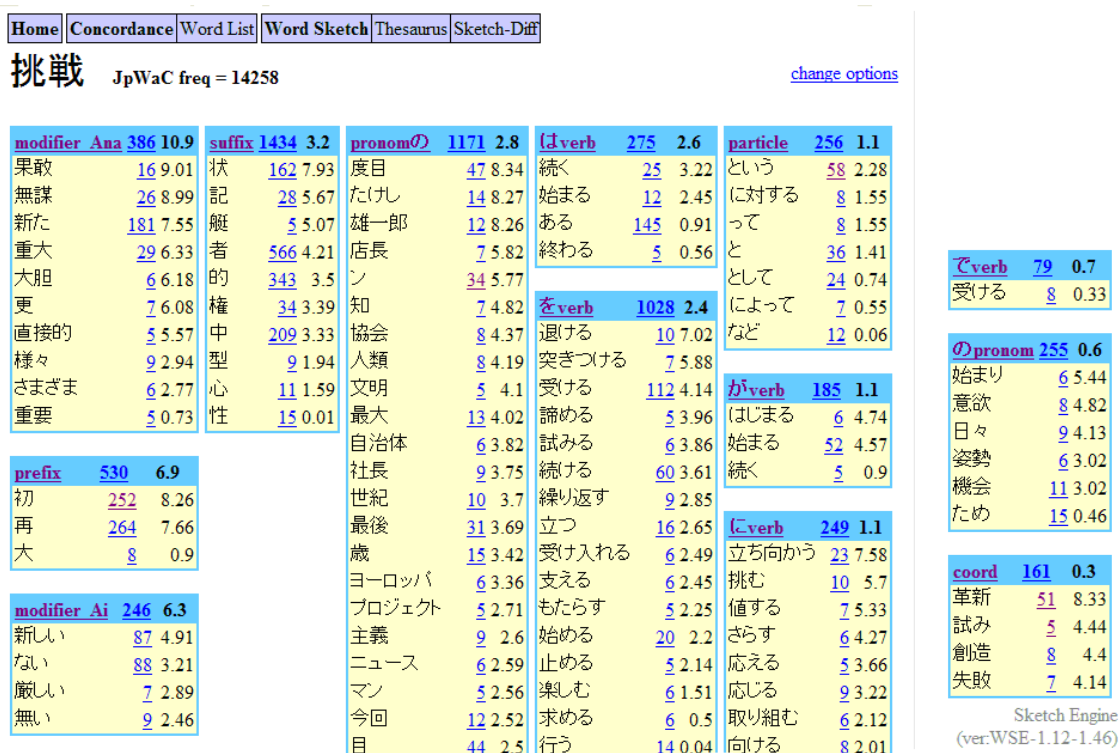


図 9 「挑戦」の Word Sketch 結果

図 9 の最左欄の「modifier\_Ana」と「modifier\_Ai」は「挑戦」を修飾する形容動詞と形容詞を示している。同じく四番目と五番目の欄の「は verb」, 「を verb」, 「が verb」, 「に verb」は「挑戦」がどの動詞と共起しやすいかを表している。一つの文法関係にある様々な共起が意味的に区別でき、異なる文法関係の中にも出現する単語間にも意味的な関係が見られる。まず、図 9 の文法関係を一覧すると、「新しい挑戦, 新たな挑戦, 挑戦は始まる, 挑戦を始める, 挑戦が始まる, 始める」, また「挑戦を試みる・もたらす・続ける」の例は善意・積極的・前向きのイメージを持ち、英語の‘initiation’と‘trial’の意味に対応していることがわかる。これらは「自分の意志で前向きに新たな難しそうな物事にあたってみる」という意味を表現している。

さらにもう一つの積極的なイメージと結ぶ用法は「挑戦を楽しむ・求める」, または「挑戦に値する」である。

一方、「挑戦」は「再-」「失敗」「諦める」など挑戦の不成功を意味する語と共起する例もある。「厳しい挑戦」, 「無謀な挑戦」, 「挑戦を繰り返す・退ける・諦める・止める」, また、「重大な・



### 3.1.2 共起表現辞典との比較

『日本語表現活用辞典』は 2004 年に刊行された日本語の学習者のための最初の共起表現辞典である。この辞典は、「語の意味」の記述を中心としている従来の日本語の国語辞典と異なり、「語の結びつき」を示している（姫野 2004）。例文とコロケーション情報が豊富に記載されており、日本語学習者などに有益なツールとなっている。日本語の大規模な均衡コーパスが存在しない状況で編纂されたため、様々な言語資源（他の辞典、文芸作品、新聞データなど）を利用している。

このような辞典を編纂する際に Word Sketch を利用するメリットがあるかどうかを示すために、辞書の中から下記の 10 項目をランダムに選び、辞書の記述を Word Sketch の結果と比較してみる。

うつむく【俯く】、かすか【微か】、くるしむ【苦しむ】、しめる【閉める】、たべる【食べる】、とめる【泊める】、はこぶ【運ぶ】、べつべつ【別々】、めいりょう【明瞭】、わる【割る】

比較の結果を 1) 多数の文法関係のための利用、2) 共起の選択方法のための利用、3) 例文選択のための利用、4) 語彙意味論的な情報のための利用という 4 項目にまとめ、いくつかの例を挙げつつ以下で考察する。

#### 1) 多数の文法関係のための利用

『日本語表現活用辞典』の中では 1,180 語の動詞と 364 語の形容動詞についての共起項目が記述されている。動詞の項目は格助詞（「が」、「を」、「と」、「に」）と結びつく名詞と動詞を修飾する副詞に関する共起情報が、形容動詞は「な」と「の」の形で修飾する名詞と「に」「て」などの形で修飾する副詞的用法との共起情報が、それぞれ示されている。一方、Sketch Engine には文法関係が 22 あり、それぞれは 2 項関係によって示されるが、二つ以上の品詞を含む関係もあるので、辞書の項目より大分多い。まず、動詞と形容動詞の項目だけでなく、それ以外の項目（特に名詞、形容詞、副詞の共起）も検索できる。またそれぞれの品詞は多数の共起項目を含む。例えば、動詞の項目では、格助詞「が」、「を」、「と」、「に」以外に、「で」、「まで」「から」、「へ」などの格助詞、および係助詞「は」と結びつく名詞も表示する。それに加えて、副詞、非自立動詞、接尾動詞との共起、他の自立動詞との並列関係などもある。

辞書としては、スペースの観点からすべての共起項目を記載することは不可能であり、辞書編集方針に委ねられることになる。一方、Sketch Engine では共起頻度と多様で統計的な重要度が計算でき、その情報に基づいて様々な共起の種類とその中で最も重要となる共起項目を提供することができることから、辞書編纂のためには大きく貢献できると考えられる。この点では Sketch Engine の優位性が明らかだが、その一方、『日本語表現活用辞典』との比較から、Sketch Engine に欠けている点も判明した。たとえば今回、用意されていた文法関係には<形容動詞語幹+に+動詞>の共起項目が含まれていないということが明らかになった。

## 2) 共起の選択方法のための利用

次に個々の文法関係の中に現われる様々な共起について述べる。

『日本語表現活用辞典』には、一つの共起タイプの中に多数の共起の例があるが、コーパス中の頻度が高いものが必ずしも挙げられているわけではない。例えば、「かすか・微か」はコーパス中で「かすかな記憶」がかなり出現するが、辞書にはない。辞書の項目内容としては、共起が多い項目と少ない項目がある。また、例文中にしか共起表現が出てこない項目もある。これらのことから共起の頻度にもなう選択方法のために **Word Sketch** を利用すると、この種の辞書に役に立つと考えられる。

一方、辞書にある共起が **Word Sketch** の結果に見られないことがある。その理由は **Sketch Engine** では **Web** データを用い、『日本語表現活用辞典』では主に小説を用いているためと考えられる。このことを明らかにするためには、**Sketch Engine** のツールに他のコーパスを載せ、多様で大量のデータに基づいた結果を比較する必要がある。

## 3) 例文選択のための利用

辞典では共起表現を例示するために、最も重要な共起と文型を示すために厳選した例文を出さなければならない。**Word Sketch** は様々な共起表現のなかで高頻度の結果を得た上でそのような例文選択をすることに役に立つ。例えば、「うつむく」の辞書項目では典型的な例として動詞・名詞・副詞が現れている（赤い顔をし、うつむいた／しばらくうつむいて考えていた／花がうつむいている／うつむいてしまった／うつむいて歩く）。一方、**Word Sketch** のデータ（図 12）を利用すると例文の中に「はずかしそうにうつむいていた／うつむいてひっそりと泣き出した／下をうつむいたまま／無言でうつむく／花の色が濃くややうつむき加減に咲く／うつむきがちであった／ちょっとうつむきかげんの頭／少しうつむきながら」など辞書にはない表現例があり、この中からも共起の候補が考えられる。このようなデータを参照しながら検討することで、さらに辞書の例文を精選することができる。

<b>modifier Adv</b> 63 8.3	<b>coord</b> 350 4.0	<b>nounで</b> 34 2.0
やや 6 6.55	ふり向く 4 8.22	無言 4 6.22
終始 3 6.03	かげる 3 7.29	
しばらく 5 4.27	うなだれる 3 6.97	<b>nounを</b> 104 1.6
少し 10 3.91	考え込む 3 5.72	顔 21 3.18
いつも 4 3.67	微笑む 4 4.78	肩 3 2.98
ずっと 3 3.23	咲く 5 3.74	頭 5 1.32
ちょっと 7 2.99	閉じる 4 3.54	下 7 1.29
また 3 2.43	黙る 4 3.47	
	泣く 7 3.13	<b>nounに</b> 41 1.2
<b>nounは</b> 127 8.0	歩く 19 2.98	げ 3 2.79
彼女 13 2.79	座る 7 2.89	そう 9 0.34
彼 10 1.54	見つめる 3 2.53	
	疲れる 3 2.04	<b>nounが</b> 24 0.9
	流す 4 2.01	花 3 1.68
	答える 5 1.89	
	笑う 3 1.47	<b>suffix</b> 18 0.3
	上げる 5 1.17	かげん 7 9.85
	立つ 4 0.67	がち 5 3.66
	話す 4 0.66	
	生きる 4 0.34	
	<b>bound V</b> 173 3.5	
	しまう 31 0.8	

図 12 「うつむく」の Word Sketch の検索結果

#### 4) 語彙意味論的な情報のための利用

『日本語表現活用辞典』のような辞書編纂の過程において、Word Sketch 以外に他の Sketch Engine の機能、Thesaurus と Sketch Difference を用いると、さらに詳しい類義語、反対語とその差異などの語の意味的な情報が得られる。

例えば、辞書の中で項目 A とその共起が他の項目 B と類似している場合には、A の項目の中だけで共起が表示され、B 項目の共起は A 項目を参考として表示する。例えば「閉める」と「閉まる」、「泊める」と「泊まる」の項目で例示できる。Sketch Difference の機能で「閉める」と「閉まる」を調べると、「～を閉める」と「～が閉まる」の一般的な自他動詞の違い以外に、幾つかの興味深い結果が見られた。まず、「閉める」は「られる」、「させる」、「っぱなし」という接尾とよく共起している（「ドアがきちんと閉められています／雨戸を閉めさせる／カーテンを閉めっぱなしにする」など）。また、「閉める」は助詞「で」との結びつきとして「後ろ手で」、「手で」、「鍵で」が現れている。複合助詞としては、「ために」も出現する（安全のために窓を閉めている）。動詞に付く自立・非自立・接尾の動詞としては、「閉めきる、閉め直す、閉めてくれる・いただく・もらう」の様な共起がある。

一方「閉まる」を見ると、「閉まりかける、閉まっておる、閉まり始める」などの共起が見られる。よく共起する名詞「ドア、窓、カーテン、シャッター」は「閉める」と「閉まる」の場合には共通して見られるが、「レストラン」、「商店」、「図書館」は「閉まる」としか共起していない。

さらに、「閉める」・「閉まる」とともに辞典には見られなかった主な名詞・動詞との共起は「雨戸」、「扉」、「蓋」、「元栓」、「蛇口」、「バルブ」である。辞書にある「障子」と「襖」の共起が

Web コーパスには現れていないのは興味深い。それは Web データの方が現代日本語を反映しており、社会の変化とともに、変化する言語の様相を表しているとも言えるだろう。

以上の比較の結論として、Word Sketch 日本語版は共起表現辞典の編集において様々な共起情報を提供し、語の意味的な情報を追加し、例文選択を支援すると考えられる。また、共起辞典だけではなく、国語辞典、日本語学習辞典、二言語辞典、シソーラス、類義語および反対語辞典、文型辞典などの様々な種類の辞書編纂を支援すると考えられる。編集時間の面からは調査していないが、英語バージョンの実績から推測すると時間的なメリットもあるはずである。一方では、日本語共起辞書を含む様々な資源を利用することにより、Sketch Engine の文法関係ファイルがさらに向上できることがわかった。

### 3.2. 言語研究と Sketch Engine

上記で検討した辞書学分野以外に、他の言語学研究でもコーパス資源の利用・方法の開発が広がっている。Sketch Engine はこのような言語学における実証的な研究方法のために利用できる。Word Sketch, Thesaurus, Sketch Difference の結果は主に語彙意味論的なデータに集中しているが、コーパス言語学の面からもそれ以外の興味深いデータが見られる。本章ではこのような構造的な情報についていくつかの例を挙げ、また、Concordance 機能によって、どのように様々な言語学的な課題（例えば文法文型など）が複数の単位として検索できるかを述べる。

#### 1) 形態論的な派生・屈折

形態論的な派生の項目を次に示す。

- suffix (接尾辞), prefix (接頭辞)
- suffix\_base (接尾辞が付く語幹), prefix\_base (接頭辞が付く語幹)
- bound\_V (キーワードの動詞に付く自立・非自立・接尾の動詞)
- V\_bound (キーワードの自立・非自立・接尾の動詞はどの動詞によく付くかを示す)

例えば、suffix の例として動詞の「聞く」の接頭辞は「お」（お聞きしたい）、bound\_V

「取り組む」、V\_bound 「みる」、「くれる」、「くださる」など（聞いてみる、聞いてくれないか）がある。また、「本」を検索すると接頭辞として現われる例（本研究、本サービス）があり、接頭辞と一緒に現れる名詞としての例もある（ご本、御本、各本など助数詞）。さらにそれらの語に後接する助数詞も共起結果の中で確認できる（～冊の本）。

Sketch Difference 機能で接尾辞「性」・「さ」とその語幹との共起の差異を確認できる。「性」は「可能」、「方向」、「生産」などの形容動詞・名詞の語幹に付き、「さ」は「大きい」、「高い」、「長い」などの形容詞の語幹に付くことなどは初級段階の学習によって獲得されるが、上級レベルに進む段階では共起しない語形を知ることが重要になる。形容動詞でも「豊か」は「性」を付けず、「便利」は「さ」をよく取る。また両方とも共起できる語（「正確さ/性」、「複雑さ/性」）もあることで、使用に当たって学習者がしばしば迷うところである。このような学習者にわかりに

くい組み合わせ・表現も素早く確認できる。

形態論的な屈折のデータの中で名詞と格の組み合わせ、または動詞に付く動詞接尾（例えば受身文や使役文など）と共に起る助動詞の情報は学習者にとって辞書などからは得にくいですが、Word Sketch を利用することにより情報が得やすくなる。また、動詞・形容動詞・形容詞は Word Sketch の結果では、lemma として表示されるため、これらの活用形の情報も直接抽出できることが期待される。

## 2) パターンを探す

共起の情報の中では、動詞の目的語は何であるか、主語あるいは話題は何であるかなどの文法関係も現れる。さらに Concordance 機能を用いると様々な簡単なあるいは複雑なパターンが探せる。Concordance での検索方法についてはすでに 2.2 と 3.3.1 に述べた。ここでは、Concordance の CQL 機能を利用して、どのように検索できるかについていくつかの例を挙げる。

### ● 「～から／で作られる」文型表現の例

「～から作られる」と「～で作られる」という文型表現の類似した用例、その共通点・差異を Concordance で参照できる。それぞれのパターンを検索するために、CQL ボックスに以下のように文字列を入力する。

```
[word="から"][word="作ら"][lemma="れる"]
```

```
[word="で"][word="作ら"][lemma="れる"]
```

「れる」を lemma とするのは終止形、連体形のほか、未然形、連用形などのような活用形を含むためである。頻度の結果としては、「から」の表現は 432 回であり、一方「で」の表現は 2,975 回現れている。それぞれの結果を Collocation candidates（コロケーション候補）機能で探すと以下のことが見られる。

「から」は以下の語と共に起している。

- 天然の材料・部品（「植物，米，木，ブドウ，サトウキビ，素材，パルプ，食材，原料，古紙，土壌，石油」）
- 抽象的・内容的な部分（「言葉，反省，意味」）
- 時代や考え方の視点（「視点，時代，頃，年，以降，観点」）

「で」の表現は以下の語と共に起している。

- 場所・作成者（「工場，＜場所＞＋の中，日本，アメリカ，＜場所＞＋の上，ドイツ，中国，国，体内，家庭，地域」など），作成者（「自分，スタッフ」など）
- 目標（「目標，要素，狙い，つもり」），考え（「コンセプト，発想，イメージ，考え」など），方法（「技術，方法，前提，技法，協力，手順，主導，人工」），状況（「段階，レベル；短期間；コスト，予算」など）
- 材料（「木，素材，材，材料，石，紙，物質，金属，葉，ガラス，木材，竹，合金，金，食材，ビーズ，プラスチック，卵，大理石」など）

材料の共起に限定して見てみると、「から」「で」とともに共起する「木，食材」などがあるが、



「から」は天然のもの・原料，一方「で」は人間が作ったもの・物質から作られているものと共起する傾向が見られる。

● 使役の「動詞+させてあげる」文型表現の例

さらにもう一つの例として使役の「動詞+させてあげる」文型表現の探し方を示す。まず Concordance の CQL ボックスに [tag="V.\*"][word="せ|させ"][word="て"][lemma="あげる"] と入力する。これによって，どの種類の動詞のあとでも，使役の助動詞「せ」あるいは「させ」が接続し，さらに「て」に続く補助動詞「あげる」のすべての活用形の連なりからなる表現を探ることができる。この形式で検索すると，Web コーパス頻度は 1,170 となる。この形式の中の高頻度動詞を調べるために CQL の[word="せ|させ"][word="て"][lemma="あげる"]の結果を Collocation candidates の機能でさらに検索すると，10 回以上現れている動詞の語幹は次のものとなる。

す (サ変動詞)，食べ，聞 (き)，休 (やす)，や (やるの語幹)，持 (も)，喜 (よろこ)，知 (し)，飲 (の)，会 (あ)，樂し，行 (い)，読 (よ)，気づ，遊 (あそ)，勝 (か)，甘え

● 複合語「気にする」の例

複合語の共起も Concordance で検索できる。例えば，[word="気"][word="に"][lemma="する"]を検索すると 10,845 の例文が出てくる。Collocation candidates 機能を使うとそのうちに「～を+気にする」の例文が 4,000 例近くある。一番多く現れる目的語を図 13 の語彙成分セット (lexical sets)として示す。

語彙成分セット	共起
他人の意見・評価・予言に関して	他人，世間体，反応，評価，近所，回り，周り，評判，世論，人，周囲，思惑，目線，眼，目，視線，外見，占い，迷信
自分・人々の姿・様子に関して	ダイエット，スタイル，体裁，年齢，ファッション，服装，汚れ，髪型，体重，太り，見た目，容姿，見栄え，日焼け
天気に関して	雨，天気，予報，天候，紫外線
病気・痛みに関して	病，痛み，具合，傷，持病，健康
財政に関して	予算，株価，料金，値段，財布，燃費
時間に関して	時間，時計，時差，終電，遅れ
知覚刺激	匂い，騒音，煙，臭い，雑音
結果に関して	成績，点数，結果，間違い，敗戦
中身の質	中身，歌詞，文法，細部，画質，音程
その他 (順番・量・距離等)	順位，順番；数値，率，数，容量；距離；動向など

図 13 「気にする」の目的語として共起する語の語彙成分セット

さらに、「気にする」の前に来る語の語彙成分セットとして「～のことを気にする」を調べると、`[word="の"] [word="こと|事"] [word="を"] [word="気"] [word="に"] [lemma="する"]`から「人間」を表す語が多いことがわかる。例えば、「子供、彼、あなた、人、私、～さん、～者」などである。

上記に示した目的語の語彙成分セットの多くは、近代の小説コーパスを利用した研究結果 (Srdanović 2007) でも観察されている。

コーパスの中でマークアップされているデータが文法関係ファイルに適切なパターンとして設定されていれば、Word Sketch などの機能によって、容易に質の良い結果が得られる。コーパスデータのマークアップとしてさらに様々な文型表現、複合用語、モダリティなどを追加すれば、Word Sketch の結果もそれに応じてさらに言語学的な情報が豊かになると考えられる。

### 3.3. 第二言語学習と Sketch Engine

第二言語学習にも近年、コンピュータ資源が利用されるようになってきている。Sketch Engine の多数の言語版はそのためにも使われており、その有用性が検討されている。日本語版も国内と海外で拡大している日本語学習に役立つ資源になり、日本語教授者と日本語学習者を支援することが可能となるであろう。ここではいくつかの面から Sketch Engine の応用の可能性を考える。

#### 1) 利用者の面

まず、日本語を第二言語として学習する場合の Sketch Engine の応用の可能性を利用者の側から見る。大きく分けると (a) 日本語の教師と (b) 日本語の学習者であるが、さらに以下のように分けられる。日本語の教師には日本語母語話者と非母語話者があり、両方とも Sketch Engine を難なく使えるはずである。しかしながら、教師といえども非母語話者である場合は、例文に現われる微妙な言語的な問題に関しては簡単に判断できない時もあり得る。一方、学習者の場合には日本語の能力レベルによって Sketch Engine の利用方法を考えるべきである。中級学習者および上級学習者はさほど問題なく使えるが、初級学習者は直接利用するのが難しいと考えられる。

さらに、学習者を利用者として考えると、ツールを利用しながら言語を学ぶ上でいくつかの難しい点も見られる。その一つは分からない語あるいは読めない漢字が現れる場合である。また、ツールの中で学習能力レベルの情報を含んでいないことである。現在のコンピュータ支援システムの中には、レベルを考慮して作成されているシステムがある。一例として、学習例文を選択するために能力レベル情報を利用する「なつめ」という作文学習支援システムがある (Nishina & Yoshihashi 2007)。このシステムは多数のコーパスを利用して、学習者の日本語能力レベルに合わせた例文や共起表現が検索できる。コンコーダンスラインでも例文の複雑さによってソートできるという研究も行われているので (Smrž 2004)、将来的にこのような機能を Sketch Engine に追加すれば、このツールは学習者のためにさらに使いやすくなると考えられる。また、日本語版では、必要に応じて振り仮名も見られるようにすることが考えられる。

## 2) 言語の四技能の面

「読む」「書く」「聞く」「話す」という言語の四技能の面から見ると、**Sketch Engine** は書くスキルの向上を支援することが第一の目的と言える。同時に、コーパス例を読むことで間接的に読むスキルの向上も支援できる。

### 3) 学習目的の面

ツールの応用は学習目的によっても分けられる。

- (a) それぞれの言語学的な知識を習得するため(例えば、語彙意味的な知識、文法パターンなど)
- (b) 言語の能力を評価するため (テスト作成)
- (c) 教科書などの学習資源を作成するため
- (d) コンピュータ学習支援システムを構築するため

一方、学習内容としてのそれぞれの言語的な知識としては、語彙意味論的な問題、形態的な問題、文法用例などについてすでに述べた (3.1 と 3.2)。従来の **Sketch Engine** における教育利用の経験では (Smrž 2004), 共起表現だけではなく、**Sketch Difference** 機能で得られる類義語の異同に関する比較情報の利用が第二言語教育において最も役に立ったと述べられている。**Thesaurus** を利用すると、特定の意味を表す語彙に関しての知識が評価できる。学習者が意味的に関連している単語の差異を知っているかどうかをチェックするためのテストが **Sketch Engine** ツールの上で自動的に作成できると Smrž (2004) は述べている。教科書などの学習資源の作成では前述した **Sketch Engine** から得られる様々な言語学的な情報が利用できる (語彙・形態素・文型パターンの振舞いや典型的な例文などである)。さらにコーパスから学習資源作成のために利用できる語彙リストが得られ、そのリストから高頻度語彙について特別に学習者の注意を向けさせることができる。**Sketch Engine** で得られた結果は高頻度の共起・類義語・反対語を示す機能を持つコンピュータ支援システムに発展させることが可能であり、例えば前述の「なつめ」システムの構築とその向上にも利用できる。

### 4) 利用場所・時間

さらに、どこでツールを利用するかによって、次のような三つの可能な方法が考えられる。

- (a) 授業時 (コンピュータ教室)
- (b) 授業の前後 (授業のための準備, 宿題など)
- (c) 遠隔教育

**Sketch Engine** のいくつかの言語版はすでに第二言語学習に利用されている。中国語学習者には中国語の **Sketch Engine** がどの程度役立つかを検討するための調査も行われている (Smith et al. 2007)。日本語版でも言語学習における可能な利用を検討するためにはこのような調査を行う必要がある。

### 3.4. その他の利用方法

上述の Sketch Engine の応用以外にも他の目的で、他の分野の研究における利用が考えられる。社会言語学の面から見ると、例えば 2.3 で「女の子」と「男の子」の例で見たように様々な興味深い現象が検討できる。単語とその共起を通じて社会通念、様々な固定観念の型が見られ、現在の大規模 Web コーパスのデータを用いると、文化研究などの分野の面から Web メディアも分析できる。さらに、多種のコーパスを比較すれば、各々のコーパスの特徴も検討できる。大規模コーパスデータの結果と人手による意識調査で得られた結果を比べるのも興味深い。例えば、アンケートによって得られた単語と単語の連想を表しているワードマップを Word Sketch と Thesaurus の結果と比較することもできる (Joice 2005)。Sketch Engine の結果には、コーパスが分析されている形態素分析ツール ChaSen の様々なタイプの誤りが発見できるため、このようなツールをテストし、ChaSen の解析精度の向上に役立てることも可能である。

ユーザーも Corpus Builder 機能を利用して、自分で他のコーパスを Sketch Engine に載せることができる。また、WebBootCat を利用して、Web から専門分野などの特定領域のコーパスを構築することもできる (Baroni et al. 2006)。

## 4. まとめと今後の課題

日本語版の Sketch Engine の最初のバージョンが作成された。本稿ではまず日本語版の作成ステップについて以下のことを述べた。

- 1) ChaSen で分析した 4 億語の Web コーパスを載せたこと
- 2) 正規表現と ChaSen の品詞から日本語の文法的関係ファイルを作成したこと

次に、伝統的なコーパス検索方法と比較することで Sketch Engine の利点を示し、大規模なコーパスから短時間で言語学的な情報を取り出すための新しい方法を紹介した。ツールのもつ様々な機能 Word Sketch, Thesaurus, Sketch Difference, Concordance は、コーパス辞書学において、有効で良質な語彙意味論的なデータが得られることから、新しいコーパスの検索手法・利用方法として役立つという結論に至った。その他の様々な分野として言語学、第二言語学習等の応用も例を挙げて検討した。

今後の課題としては日本語版の以下の問題点を改良し向上させる予定である。

- 1) Web コーパスのデータをさらにクリーンにし、データのテキスト分類情報を含む様々なメタデータを追加する。
- 2) 複合語を含む文法的な関係をさらに追加・向上させ、他の形態素解析ツールのメリットも組み込む。
- 3) 母語話者・学習者向けの追加情報を付与する。日本語での ChaSen タグ、日本語での文法関係名称、ChaSen による仮名、ローマ字、読みを表示する機能を追加する。

## 参考文献

- Srdanović Erjavec, Irena (2007) 「コーパス辞書学における意味分析－「気にする」と「気にかける」の慣用句を例として」『日本語教育連絡会議論文集 19』, 83-89, 日本語教育連絡会議事務局
- 投野由紀夫 (2007) 「日本語コーパスでの Sketch Engine 実装の試み」『特定領域研究「日本語コーパス」平成 18 年度公開ワークショップ (研究成果報告会) 予稿集』, 109-112, 文部科学省科学研究費特定領域研究「日本語コーパス」総括班
- 姫野昌子 (2004) 『日本語表現活用辞典』 研究社
- Baroni, Marko, Adam Kilgarriff, Jan Pomikalek & Pavel Rychly (2006) WebBootCaT: a web tool for instant corpora, *Proceedings of the EuraLex Conference 2006*, 123-132.
- Baroni, Marko & Silvia Bernardini, eds. (2006) *Wacky! Working papers on the Web as Corpus*, Bologna: GEDIT.
- Church, Kenneth Ward & Patrick Hanks (1989) Word association norms, mutual information, and lexicography, *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, 76-83.
- Crystal, David (2006) *Language and the Internet*, Cambridge: Cambridge University Press.
- Erjavec, Tomaž, Kristina Hmeljak Sangawa & Irena Srdanović Erjavec (2006) jaSlo, A Japanese-Slovene Learners' Dictionary: Methods for Dictionary Enhancement, *Proceedings of the 12th EURALEX International Congress*
- Erjavec, Tomaž, Adam Kilgarriff & Irena Srdanović Erjavec (2007) A large public-access Japanese corpus and its query tool, *CoJaS 2007, The Inaugural Workshop on Computational Japanese Studies*.
- Gahl, Susanne (1998) Automatic Extraction of subcategorization frames for corpus-based dictionary-building, *Proc EURALEX 1998*, 445-452.
- Ghani, Rayid, Rosie Jones & Dunja Mladenic (2001) Using the Web to Create Minority Language Corpora, *Proceedings of the 2001 ACM CIKM: Tenth International Conference on Information and Knowledge Management*, 279-286.
- Heid, Ulrich, Stefan Evert, Vincent Docherty, Wolfgang Worsch & Wermke, Matthias (2000) Computational tools for semi-automatic corpus-based updating of dictionaries, *EURALEX 2000 Proceedings*, 183-196.
- Joyce, Terry (2005) Constructing a large-scale database of Japanese word associations, In Katsuo Tamaoka (ed.) *Corpus Studies on Japanese Kanji (Glottometrics 10)*, 82-98, Tokyo: Hituzi Syobo & Germany: RAM-Verlag:Ludenschied.
- Keller, Frank & Maria Lapata (2003) Using the Web to Obtain Frequencies for Unseen Bigrams, *Computational Linguistics* 29 (3), 459-484.

- Kilgarriff, Adam & Michael Rundell (2002) Lexical Profiling Software and its Lexicographic Applications - a Case Study, *EURALEX 2002 Proceedings*, 807-818.
- Kilgarriff, Adam, Pavel Rychly, Pavel Smrž & David Tugwell (2004) The Sketch Engine, *Proc. Euralex*, 105-116.
- Kilgarriff Adam & David Tugwell (2001) WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography, *Proc. workshop "COLLOCATION: Computational Extraction, Analysis and Exploitation. 39th ACL & 10th EACL*, 32-38.
- Nishina, Kikuko & Kenji Yoshihashi (2007) Japanese Composition Support System Displaying Occurrences and Example Sentences, *Symposium on Large-scale Knowledge Resources (LKR2007)*, 119-122.
- Rundell, Michael, ed. (2002) *Macmillan English Dictionary for Advanced Learners*, London: Macmillan.
- Sharoff, Serge (2006) Open-source corpora: using the net to fish for linguistic data, *International Journal of Corpus Linguistics* 11(4), 435-462.
- Smith, Simon, Alice Chen & Adam Kilgarriff (2007) A corpus query tool for SLA: learning Mandarin with the help of Sketch Engine, *Practical Applications in Language and Computers - PALC 2007*
- Smrž, Pavel (2004) Integrating Natural Language Processing into E-learning — A Case of Czech, *Proceedings of the Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning, COLING 2004*. 106-111.
- Srdanović Erjavec, Irena, Tomaž Erjavec & Adam Kilgarriff (2008 予) A web corpus and word-sketches for Japanese, 『自然言語処理』, 言語処理学会
- Ueyama Motoko & Marko Baroni (2005) Automated construction and evaluation of a Japanese web-based reference corpus, *Proceedings of Corpus Linguistics 2005*.

## **Sketch Engine : corpus query tool for Japanese and its possible applications**

SRDANOVIĆ ERJAVEC Irena, NISHINA Kikuko  
(Tokyo Institute of Technology)

### **Keywords**

Sketch Engine, corpus linguistics, lexicography, second language learning, collocations

### **Abstract**

Although corpus-based language research has been developing rapidly in recent years, there is still a lack of resources in regards to their size, textual variety, and time of creation, and of efficient and user-friendly corpus query tools. This is also the case for the Japanese corpus linguistics, which is one of the primary reasons for the recent rise in projects constructing Japanese corpora resources.

In this paper, we present a method for extracting linguistic information from corpora using the Sketch Engine corpus query tool, which has recently been extended for the Japanese language. The Japanese version is based on a 400 million word Japanese Web corpus, which is linguistically annotated by the morphological analyzer ChaSen, and a Japanese grammatical relations file. The tool offers efficient and user-friendly ways of extracting concise linguistic data about words—their grammatical and collocational behavior, as well as thesaurus-like information and differences in usage for similar words. We explain, through examples, how the tool could be utilized in corpus lexicography, linguistic research and computer assisted language learning of the Japanese language. The investigation part of the article concentrates mainly on the ways that the tool could be applied within the dictionary creation process, and the results illustrate how each of the tool functions can greatly contribute to that process.