日本語教育における Sketch Engine の応用

スルダノヴィッチ・イレーナ(東京工業大学)、仁科喜久子(東京工業大学)
irena_srdanovic@hotmail.com、 nishina.k.aa@m.titech.ac.jp

1　はじめに―コーパスと日本語教育

コーパスとは、コンピュータによる検索が可能な大量の言語データのことであり、「電子化テキストの集合体」とも定義されている。コーパスは現れたころから言語研究に大きな影響を与え、英語を中心に言語教育にも利用されるようになってきた。コーパスの利用は、言語教育に二つの大きな変化をもたらした。一つは言語の新たな記述方法を可能にしたこと、もう一つは主観的な判断に頼らない教育資料の作成を可能にしたことである。たとえば、学習辞書や教科書の作成、シラバス作成の新しい方法、新しい学習方法などの応用である（Hunston 2002)。日本語教育においても、こういった応用が考えられており、国立国語研究所で行われている「言の葉」プロジェクトの中で、日本語教育へのコーパスの応用に向けた活動が行われている。

言語教育における Data-driven learning では、学生達がコーパスを利用しながら、言語の用法を自ら発見することで学いんでいく（Johns 1991)。たとえば、学習者は学習文法書の記述をコーパスの結果によって評価したり、批判したりする。Data-driven learning の学習方法はさまざま存在するが、教師が学生にタスクを与えることで誘導的に学ばせることもできる。

シラバス作成、特に語彙シラバス（lexical syllabus、Sinclair and Renouf 1988、 Willis 1990）においてもコーパスは重要な役割を果たす。このような語彙シラバスは、高頻度の語彙が中心となっているが、語彙以外の項目（用法のパターン）も含まれている。

コーパスは、英語を中心とした辞書と文法書の編集において、大きな革命をもたらした（Sinclair 1987; Baught et al 1996)。すべての大手出版社は、辞書をコーパスに基づいて編集したと唱っている。伝統的な方法で作成された辞典と比べると、いくつかの点でコーパスを利用した方法のものが優れている。具体的には、辞書における項目の頻度(frequency)、共起・慣用語 (collocations & phraseology)、様々なジャンルの特徴(variations)、文法パターンにおける語彙(lexis in grammar)、実際の使用例（authenticity）に関しての情報である（Hunston 2002)。

本稿では、コーパス検索システム Sketch Engine を紹介し、その日本語教育への応用の可能性について述べる。特に、日本語学習辞典の作成への応用について検討する。

2　Sketch Engine の紹介

コーパスを言語教育に応用するためには、コーパスだけでなく、コーパスの検索を可能とするツールも必要である。その一つとして、Web 上で利用可能な Sketch Engine（SkE）というコーパス検索ツールがある。初期の SkE（Kilgarriff et al. 2004）は英語のために作成されたが、その後、他の言語のバージョンが追加された。日本語版も去年作成され（Srdanovic et al 2008)、JpWaC（Erjavec et al 2007）という 4 億語からなる大規模なウェブコーパスを有している。SkE が標準的なコーパス検索ツールと異なる点は、語に付随する文法とコロケーション情報を提示することができる「Word Sketch」という機能を持ち、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機能（「Thesaurus」

と「Sketch Difference」）も備えていることである。SkE の利用目的としては、日本語辞典編集、日本語学研究、日本語教育などが考えられる（Srdanovic・仁科 2008）。[1]

Word Sketch で単語を調べると、大規模なコーパスから得られる単語の共起関係・文法関係が表示される。「先生」という単語を Word Sketch で調べた結果の一部を図 1 に示した。左側の「が verb」の欄は、助詞「が」を伴って共起する動詞を示しており、代表的な共起としては、「おっしゃる」、「いらっしゃる」、「亡くなる」が現れることがわかる。2 列の数字については、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、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。

図 1　「先生」を Word Sketch で検索した結果の一部

（この表示では、それぞれの関係の共起項目は 10 項目までに制限されている）

3　日本語教育における SkE の応用の可能性

Sketch Engine は、英語、中国語、チェック語などの言語において第二言語教育のために利用されており（Smith et al. 2007、 Smrž 2004 など）、その有用性が検討されている。日本語版も日本語学習に役立つ資源になるであろう。

SkE の利用者については、(a) 日本語の教師、(b) 日本語の学習者、(c) 日本語の学習資源作成者

[1] ウェブデータは教育に使えるかどうか疑問に思われるかもしれないが、大規模データであるため、高頻度の項目は信頼性が高く、他のコーパスより偏っていないデータであることが示されている（Srdanović・仁科 2008）。

が考えられる。教師は、日本語母語話者でも非母語話者でもSkEが使えるはずだが、非母語話者は微妙な言語的な問題に関して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、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、初級学習者には直接利用するのが難しいと考えられる。学習資源作成者は、教師と同じように、ツールを問題なく使えるはずである。学習者が利用することを考えると、SkEにはいくつか改善すべき点が残されている。たとえば、学習能力レベルの情報が含まれていないこと、分からない語や読めない漢字が現れることなどである。

「読む」「書く」「聞く」「話す」という言語の四技能の面から見ると、SkE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、コーパス中の例文を読むことで間接的に読むスキルの向上も支援できる。

学習目的の面から見ると、SkEは(a)言語学的な知識を習得するため(語彙の意味的な知識、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・学習辞典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などに利用できる。特に、学習辞典への応用に関しては、『マクミラン英語学習辞典』(Rundell ed. 2002)の編纂に、SkEとBNC(British National Corpus)が利用されたという実績がある(Kilgarriff & Rundell 2002)。日本語版も、本稿の4章と5章で検討するように、今後日本語の学習辞典編纂に応用できると考えられる。

さらに、利用場所・時間としては、(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)遠隔教育が考えられる。

4 Sketch Engineと学習共起辞典との比較(評価1)

本章では、日本語学習辞典の作成のためSketch Engineの応用を考え、SkEで得られた共起関係と『日本語表現活用辞典』(姫野2004)に記述されている共起関係とを比較する。『日本語表現活用辞典』は日本語の学習者のための最初の共起表現辞典であり、例文とコロケーション情報が豊富に記載されている。他の辞典、文芸作品、新聞データなどの言語資源を利用して編纂され、動詞1,180語と形容動詞364語がカバーされている。比較に用いた表現は『日本語表現活用辞典』の中からランダムに10項目[2]を選んだ。結果は以下のようにまとめることができる。

- 『日本語表現活用辞典』よりSkEの方が共起関係の種類が多い。

  『日本語表現活用辞典』の中には、動詞の共起関係として、名詞+「が/を/と/に」+動詞の共起があり、形容動詞の共起関係として、形容動詞+「な/の」+名詞、形容動詞+「に/て」+用言の共起がある。一方、SkEには文法・共起関係が22あり、それぞれが2項関係によって示され、二つ以上の品詞を含む関係もある。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関係を含む。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。
- スペースの観点からすべての共起関係とその項目を辞書に記載することは不可能であるが、SkEでは共起頻度と統計的な重要度が計算できるため、最も重要な共起関係・項目を選ぶことができ

[2] 比較した項目: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】。比較の結果の詳細については、Srdanovic・仁科(2008)を参考されたい。

るのは大きなメリットだと考えられる。

『日本語表現活用辞典』には、一つの共起関係の種類の中に多数の共起項目があるが、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。辞書の項目には、共起頻度が高い項目と低い項目がある。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、共起の頻度にと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。その理由を明らかにするためには、SkE 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。

- 辞書の中に共起表現を例示するために、最も重要な共起と文型厳選した上で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、Word Sketch で得られる「うつむく」の表現例の中には、辞書項目では典型的な例の中で現れていないものもある。「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／無言でうつむく／花の色が濃くややうつむき加減に咲く」などである。

- 『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 の機能がある。Thesaurus と Sketch Difference を用いると、類義語、反対語、およびそれらとの違いなど、さらに詳しい語の意味的な情報が得られる。

例えば、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る。「閉める」は「手で閉める」、「鍵で閉める」、「閉めきる、閉め直す、閉めてくれる・いただく・もらう」、「閉めっぱなし」などの共起がある。「閉まる」は「閉まりかける、閉まっておる、閉まり始める」などとよく共起しており、特別な共起の違いが見られる。

5　共起辞典編集への Sketch Engine の応用　（SketchEval、評価２）

本章では、大規模な日本語共起辞典の編集へ応用したときの SketchEngine の評価について述べる。この評価は「SketchEval」という多言語の SkE の評価を目指すプロジェクトの中で進行中のものである。それぞれのコーパスからランダムに取り出した 42 項目の名詞、形容詞、動詞に対する SkE の共起表現の抽出結果が、大規模な共起辞典のために適切かどうかを評価する。この評価のために、まず、品詞ごとに高頻度、中頻度、低頻度の単語のサンプルリストを作成する。サンプルリストの中に、形態素解析などの問題で評価から除外すべき単語があれば、予備リスト（reserve list）の中から適切な項目を選び、取り換える。図 2 は、日本語のために JpWaC から取り出した単語のサンプルリストと予備リストを示している。評価から除外される項目は取り消し線でマークされ、代わりに評価に利用した項目が強調されている。例えば、形容詞の中の「っぽい」という項目は接尾辞なので、その代わりに予備リストから「重い」という形容詞を選び、評価のために利用した。

図２JpWaC からランダムに取り出した単語のサンプルリストと予備リスト

（名詞は名詞-サ変接続（N.Vs）と名詞-形容動詞語幹（N.Ana）を含んでいる[3]）

source: jpWaC

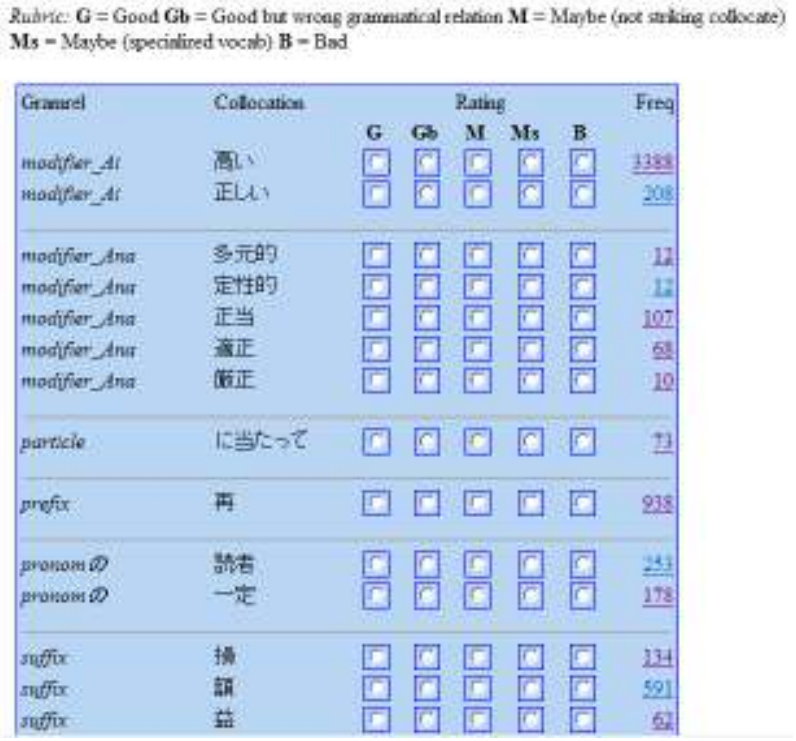
	Sample lists			Reserve lists		
	Nouns	Verbs	Adjectives	Nouns	Verbs	Adjectives
Common (100-2999) minfr = 10067 maxfr = 481866	N = 1963 急 研究 完成 男性 緑 評価	N = 367 生まれる 扱う 支払う 忘れる	N = 74 よろしい ~~っぽい~~ 素晴らしい 大きい	心配 箱 積極 プロセス 地区 建設	ちゃう 知れる 語る 受け入れる	重い 長い 忙しい 無い
Mid (3000-9999) minfr = 1931 maxfr = 10061	N = 5326 欠席 蓄積 マスター 俳句 情勢 有力	N = 840 まつ ~~文~~ 資する 溜まる	N = 105 黒い おとなしい こい 親しい	フォント 刑事 澄 蝶 包装 メス	溢れる 拒む 隠る しむ	柔らかい むずかしい きつい とんでもない
Low (10,000-30,000) minfr = 324 maxfr = 1930	N = 15076 ~~クレイ~~ 方角 近鉄 走り ~~苑~~ 人妻	N = 2114 駆け込む やせる 書き留める 滅ぶ	N = 299 むつかしい 仲良い くすい 腹立たしい	水槽 グローバリゼーション 青島 懇話 射精 モグラ	対する 振舞う 吸い上げる くぼむ	若々しい 面倒くさい 耐え難い 香ばしい

図３は、SketchEval のページとして、SkE が抽出した「評価」という単語の共起項目を示している。評価する項目は各語につき 20 共起ずつ表示される。この評価は進行中であり、10 月 9 日時点では次のような結果が見られる。89.13%の項目の共起は良い共起(Good)と判断されており、それ以外に7.17%は「共起が強くない」(Maybe (not striking collocate))、3.04%は良くない共起 (Bad) の判定となっている。現段階では、評価結果の傾向として以下のようにまとめられる。

- 高頻度の共起のほうが低頻度より評価が高い
- Maybe (not striking collocate) (共起が強くない) の評価は以下の場合によく現れる：
  - 表記のゆれ（漢字＆平仮名）(たとえば、 「目指す/めざす/目ざす」)
  - 低頻度の共起、コーパス中の同じ例文の繰り返し、一つのウェブページの中での共起の繰り返し(例文が違う)
  - 派生的な語との共起（たとえば、「完成」という項目の共起としての「楽曲の完成者」）
  - 固有名詞、地名など（「池子の緑」、「近鉄の中村・山口」）
- Bad（良くない共起）の評価は以下の場合によく現れる:
  - 形態素解析の間違い（たとえば、「よろしくぅ」は「よろしい」と「くい」に解析さて、「よろしい」の共起として「くい」が現れる。また、「急がば回れ」が「急+が+ば+回れる」と解析され、「急が回れる」という共起として現れる。）
  - コーパス中の例文の繰り返し
- 良い共起 (Good) と判断されている項目の共起が 89%程度あることから、SkE が辞典編集に有効であると認められる。

[3] JpWaC は「茶筅」という形態素解析ツールで解析されているため (http://chasen.naist.jp/stable/ipadic/)

図 3「評価」という名詞を評価するための共起項目（SketchEval 表示）



Rubric: **G** = Good **Gb** = Good but wrong grammatical relation **M** = Maybe (not striking collocate)
Ms = Maybe (specialized vocab) **B** = Bad

Gramrel	Collocation	G	Gb	M	Ms	B	Freq
modifier_Ai	高い						1388
modifier_Ai	正しい						308
modifier_Ana	多元的						12
modifier_Ana	定性的						12
modifier_Ana	正当						107
modifier_Ana	適正						68
modifier_Ana	厳正						10
particle	に当たって						73
prefix	再						935
pronom の	読者						243
pronom の	一定						178
suffix	操						134
suffix	論						591
suffix	益						62

6　まとめと今後の課題

本稿では、Sketch Engine（SkE）というコーパス検索ツールの日本語教育への応用の可能性について検討した。特に、学習辞典の編集への応用を考え、SkE の 評価を行った。その一つは、学習用共起辞典との比較評価、他の一つは多言語の SkE の評価のプロジェクト（SketchEval)の中での日本語共起辞典編集のための評価である。評価の結果によると、学習共起辞典の編集のために SkE が有効であることが確認できた。同時に、SkE の日本語版に現れる共起関係に関して改善すべき点が明らかになった。

なお、日本語教育への様々な応用を考えると、SkE を改善する方法として、以下の点が挙げられる：

- ＳｋＥに他のコーパスを載せる
- ジャンル別の検索ができるようにする
- 能力レベル別に共起・例文を表示する
- 振り仮名を入れる

参考文献

姫野昌子（2004)『日本語表現活用辞典』研究社

Baugh, S., A. Harley and S. Jellis (1996) The Role of Corpora in Compiling the Cambridge International Dictionary of English, *International Journal of Corpus Linguistics*, 1:1, pp. 39□59.

Erjavec, T., Kilgarriff, A. & Srdanović, I.E. (2007) A large public-access Japanese corpus and its query tool, *CoJaS 2007, The Inaugural Workshop on Computational Japanese Studies*.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Johns, T. F. (1991) 'Should you be Persuaded: Two Samples of Data-Driven Learning Materials'. In Johns, T. F. and King, P. (eds.) 'Classroom Concordancing' *Birmingham University English Language Research Journal* 4: 1-13.

Kilgarriff, A. & Rundell, M. (2002) Lexical Profiling Software and its Lexicographic Applications - a Case Study, *EURALEX 2002 Proceedings*, 807-818.

Kilgarriff, A., Rychly, P., Smrž, P. & Tugwell, D. (2004) The Sketch Engine, *Proc. Euralex*, 105-116.

Rundell, M, ed. (2002) *Macmillan English Dictionary for Advanced Learners*, London: Macmillan.

Sinclair, J.M. & Renouf, A. (1988) A lexical syllabus for language learning. In: R. Carter and M. Mccarthy, eds. *Vocabulary and language teaching*. Harlow: Longman.

Sinclair, J.M.(1991) *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Smith, S., Chen, A. & Kilgarriff, A. (2007) A corpus query tool for SLA: learning Mandarin with the help of Sketch Engine, *Practical Applications in Language and Computers - PALC 2007*

Smrž, P. (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, I. E., Erjavec, T. & Kilgarriff, A. (2008) A web corpus and word-sketches for Japanese, 『自然言語処理』15/2

Srdanović, E.I.& Nishina, K. (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. 59-80.

Willis, D. (1990) *The lexical syllabus*. London: Harper Collins.