

コーパス検索ソフト Simple KWIC Lister の紹介と活用事例

秋 澤 委太郎

【要旨】

本稿では、自作の OS X 用コーパス検索ソフトウェア Simple KWIC Lister の機能を紹介し、配布するとともに、日本語授業への活用例を報告する。本ソフトはオフラインのスタンドアロン・アプリケーションとして、手軽に素早く使える上に検索や表示も高速という特長を備えており、日常の学習指導のために有用である。

例えばこのツールを用いて「配慮する」という動詞の用例を調査したところ、容易に以下のことが分かった。「配慮する」という動詞には助詞の「に」を伴う場合と「を」を伴う場合の二通りの用法があり、前者の方がより多く使われる傾向にある。そして、「人や集団、あるいはその心情」を目的語にとりたいのであれば、前者の用法を選んだほうが聞き手あるいは読み手に違和感を与えにくい。

【キーワード】

授業準備、例文検索、KWIC、コーパス、「配慮する」の用法調査、OS X、Cocoa

1 はじめに

コーパスを用いた言語学・言語習得学の研究、そして言語教育の実践はすでに数多く見られる。多くのコーパスや検索・分析ツールがオンライン・オフライン、あるいは有償・無償を問わず公開されており、いずれも非常に有用であるが、それぞれに長所と短所があり、使用者の目的に合ったものを探す必要がある。

コーパスを用いて筆者が行いたいのは、

- ① 上級日本語のクラス準備のため、手軽にそして高速に単語や文型の用例を探し、保存する。
- ② この作業を、筆者が常用している OS X で済ませる。

ということである。①のためには KWIC（キーワード・イン・コンテキスト＝コーパスから特定の語句と、それが用いられている文脈とを合わせて取り出したもの）を検索・表示できるツール、つまりコンコーダンサーが非常に役に立つ。特に、オフラインで素早く簡便に使えるデスクトップ・アプリケーションの形態が望ましい。しかしこれを実際に探してみると、Windows 環境では選択肢が豊富だが、OS X で選べるものは限られている。これらの目的に適うツールが見出し難かったため、OS X のネイティブ環境である Cocoa 上で動作するソフトウェア、Simple KWIC Lister を自作するに至った。

本稿の目的は、同様の動機を持っている読者諸氏にこのソフトを紹介、配布するとともに、実際の使用例を報告することにある。

2 Simple KWIC Lister の機能と使用例

Simple KWIC Lister は、コーパスから KWIC を抽出できるアプリケーションである。ダウンロードは下のリンクから行うことができる（.dmg 形式のディスクイメージ）。

<http://www.iucjapan.org/akizawa/SimpleKWICLister.dmg>

利用は無料である。ダウンロードされた SimpleKWICLister.dmg をダブルクリックして開き、そのウィンドウの指示に従ってアプリケーションをインストールする²⁾。

本章では、Simple KWIC Lister の機能と使用例を紹介しながらその特長を明らかにする。詳しい使用方法や対応システムは、dmg ファイルに同梱されているマニュアルを参照のこと。

図1：Simple KWIC Lister のスクリーンショット



2-1 基本的な機能の紹介

本節では、主要な 8 機能のうち操作の基本となる 5 つを紹介する。

- ① OS X のファイルシステムからアクセスできる任意のファイルやフォルダ（複数選択可）をコーパスとして用いることができる。

利用可能なファイル形式は、テキストファイル（.txt）、リッチテキストファイル（.rtf, .rtfd）、html ファイル（.html, .htm）、Adobe PDF ファイル（.pdf）、Microsoft Word ファイル（.doc, .docx）である。テキストファイルは、UTF-8 はもちろん、Shift_JIS をはじめとした各種日本語エンコーディングに対応している。ファイルのエンコーディングは自動的にソフトウェアが判定するため、利用者が自ら指定する必要はない。

本ソフトウェアの利点の一つは、コーパスを利用するにあたって品詞分解などの事前処理が一切不要なことである。分析のためのデータベースとしてコーパスを高度に活用するためには、解析ソフトに読み込ませる前にこのような下準備が不可欠だが、本ソフトは単純に例文を検索することが目的なので、文書ファイルをそのまま読み込ませている。より詳しい分析を行う必要があれば、他のソフトウェアを活用されたい³。

- ② コーパスに含まれる任意の文字列の用例を KWIC 形式で抽出、表示することができる。

近年の CPU や OS の進化を活かした並列処理により、高速に抽出を行うことができる。例えば、『新潮文庫の百冊』、CASTEL/J テキストデータベース、そして「青空文庫」から無作為に選んだ作品を合わせた全 1,575 個、計約 81.4 メガバイトのテキストファイルを「構造」という単語で検索すると、5.19 秒で 1,105 の用例が抽出された⁴。

日本語の授業のために単語の用例を探すにはこのコーパスで十分であるが、より大規模なデータを用いることももちろん可能である。例えば、日本語 Wikipedia の箇条書き部分を除く全テキストデータを 1 メガバイト単位に分割してコーパス化した、全 1,655 個、計 1.74 ギガバイトのテキストファイル⁵を同じ「構造」という単語で検索すると、53.20 秒で 71,603 の例が列挙された（前頁図 1）⁶。

- ③ 検索語として正規表現を用いることができる。ただし OS X 10.7（Lion）以降のみ。

図 2：正規表現による検索



ある時学生から、「～病」と「～症」を漢字熟語作成時にどのように使い分けたいのか質問されたことがあった。そこで Simple KWIC Lister を使い、「[一-脛]+病」そして「[一-脛]+症」という文字列で検索し、それぞれ前頁図2のような結果を得ることができた。

「[一-脛]」という文言は、「文字コード体系 UTF-8 に含まれるすべての漢字のうち任意の1文字」という意味であり、「+」の記号は「直前の表現が一つまたは複数繰り返されたもの」という意味である。

一般的には、「～病」という表現は病気の種類を示したい時に使い、「～症」という表現は「病気が原因で身体や精神に現れた兆候」という意味で使う、と説明できるであろうが、説明者の思いついた例だけではなく現実に用いられている例を学生に示せる意義は大きい。

文法のクラスでも本ソフトが活用できる。アメリカ・カナダ大学連合日本研究センターでは初級文法の復習授業を行っているが、そこで動詞の使役形の使い方を指導する際、使役の客体がどの助詞をとるかはよく学生が混乱する問題である。例えば「彼に心配させる」とするか「彼を心配させる」とするか、という混乱が典型的だが、これについては「心配させる」を含め、「驚かせる」「怒らせる」「笑わせる」等、使役客体に何らかの感情を喚起する場合は「を」を使うという説明が一般的である。この場合も先と同じ要領で正規表現を活用し、簡単に例文を探すことができる。例えば「[一-脛]+させ」というキーで検索すると、いわゆるスル動詞の使役形の用例が豊富に見つかるので、その中から適切なもの、あるいは印象的なものを学生に授業で提示すれば理解がさらに深まるだろう。

④ 表からの KWIC の削除やコピー／ペースト、移動を自由に行うことができる。

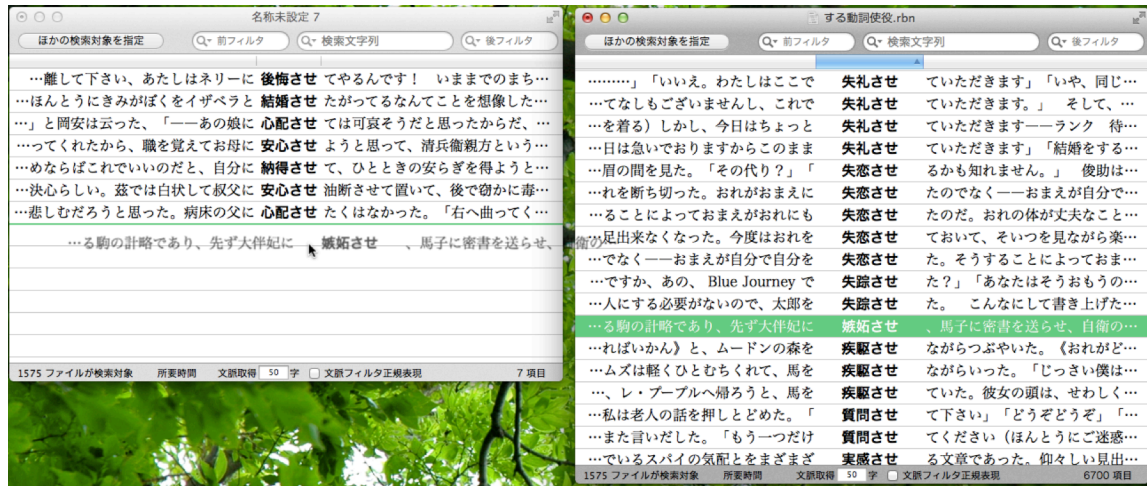
③で動詞の使役形の例を抽出したが、その表をじっくり眺めていると、例えば「あの娘に心配させては可哀想だ」（山本周五郎『さぶ』（新潮文庫、1965年）、下線は筆者による）のように、先述した「を」と「に」の使い分けの説明からは逸脱した用例がしばしば見付き、非常に興味深い。そして、興味深い例は別にまとめて取っておきたいと思うものである。このような時は、新規にウィンドウを開き、気になる例をそこにドラッグ・アンド・ドロップして追加することができる（次頁図3）。

⑤ KWIC の表を保存する、または CSV ファイルとして書き出すことができる。

検索によって得た、あるいは検索結果から自分なりに選び出して集めた KWIC の表は、.rbn 形式の独自ファイルとして保存し、将来いつでも本ソフトで開き直すことができる。ただし、②の第2段落で Wikipedia から作成した図1のような巨大データは、別途他のアプリケーションで詳しい解析を行う必要が生じるかもしれない。そのような場合は、

表を CSV 形式で書き出すことによって、例えば Microsoft Excel のようなソフトウェアで開くことができる。

図 3 : KWIC をドラッグ・アンド・ドロップで別ウィンドウに移動



2-2 発展的な機能の紹介と学習指導への活用例

本節では、「配慮する」という動詞の用法を学生に指導する際に本ソフトを活用した実例を交えながら、発展的な 3 機能を紹介する。各機能には前節からの通し番号を付した。

⑥ キーワード、前後文脈をキーとして、KWIC を 50 音順に並べ替えることができる。

この機能の使用例はすでに上掲の図 3 に含まれている。右側のウィンドウを見ると、検索語が 50 音順に整列しており、その最上部が青くなっているのが分かるだろう。つまり、この部分をクリックすると、その列をキーとして KWIC を自動的にソートすることができるのである。これは検索語の前文脈の列と後文脈の列についても同様である。文脈でソートした場合は、検索語に近い部分ほど並べ替えにおいて優先される。文脈をキーとした並べ替えは、特に動詞の統語的特徴を調べるために有用である。

先般、学生から、「配慮する」という動詞の使い方は「＜名詞＞に配慮する」か、それとも「＜名詞＞を配慮する」か、という質問を受けた。岩波書店『広辞苑』第六版を見ると「当人の気持ちを配慮する」という例文が挙がっており、三省堂『大辞林』第三版にも「相手の立場を配慮する」と例示されている⁷。しかし、筆者には「＜名詞＞を配慮する」という用法は誤りである、という感覚があった。語構成に注目すると、「配慮」は動詞「配」と直接目的語「慮」の組み合わせであり、「配」は文法的に「＜関節目的語＞に＜直接目的語＞を配する」と用いる。従って、すでに直接目的語の「慮」を含んでいる「配慮する」

は間接目的語をとって助詞「に」を用いるべきはずだ、と思ったわけである。

そこで Simple KWIC Lister を使って「配慮」の用例を検索すると 8、得られた全 235 例のうち「＜名詞＞を配慮する」が 5 件、「＜名詞＞に配慮する」は 40 件を数えた 9。

この結果を見ると、筆者の当初抱いていた、「＜名詞＞を配慮する」という使い方を誤りだとする感覚自体が誤りだったことが分かる。「＜名詞＞を配慮する」の例は、比較的少ないにせよ実在するのである。しかし一方で、「＜名詞＞に配慮する」と「＜名詞＞を配慮する」の用例を比較した時、両者の使い方には一定の区別が見出される。

「＜名詞＞に配慮する」という用法では、＜名詞＞が人や集団、あるいはその心情を意味するものとなっており、「配慮する」が「その人や集団の立場に立って思いやる」という意味で使われている場合が多い。他方、「＜名詞＞を配慮する」の用例には異なった傾向がある。ここで、その 5 例をすべて列挙しよう。いずれの例も、文字への網掛けと下線は筆者が施したものである。網掛け部分は「配慮する」の直接目的語を示している。

- ア) したがって、協働者である他の成員の働きにできるかぎり目くばり気くばりをし、全体の中の個として自分はどうふるまうのが職場システムの効果を最大にしうるかを、たえず配慮できる人間であることが求められるのである 10。
- イ) これに対して「つとめる」というのは所属する共同体の無限の発展を生活の理念として、全体の中の個としてよりよく共同体に貢献するにはどうしたらよいかをつねに配慮しつつ仕事をする事だと定義しよう 11。
- ウ) このような方向への話の展開は個人のエージングだけでなく人類のエージングを配慮するからであって、個を中心に考えて自分の一代以上のことは考えてもみようという方々にとっては無用の記述と強く感じられることであろうことは十分承知している 12。
- エ) 昼間でも雑用にわずらわされず実験ができるように、日程を配慮した 13。
- オ) ところで最近の事例を配慮して、ハドスン夫人に山嶋のえぶくろを調べるように頼んでおくかな 14

これらの例でも、文や節が全体として「誰かの立場や心情、あるいは便益を慮る」という意味を形成する傾向は強い（例ア、イ、ウ、エ）が、『広辞苑』や『大辞林』の例文とは異なり、「配慮する」の直接目的語は「人や集団、あるいはその心情」と呼べる範疇の名詞（節）ではない。

つまり、「配慮する」という動詞には「に」を伴う場合と「を」を伴う場合の二通りの用法があり、前者の方がより多く使われる傾向にある。そして、「人や集団、あるいはその心情」を目的語にとりたいのであれば、前者の用法を選んだほうが聞き手あるいは読み手に違和感を与えにくい、ということだ。これで筆者が当初抱いていた「＜名詞＞を配慮

する」に対する違和感の正体が明確になり、より適切な説明を学生に提供できるようになった。

ところで、Simple KWIC Lister は検索語とその前後それぞれ 50 文字（字数は 200 文字までの範囲で任意に変更可能）を抽出することから、KWIC は文単位にならない、つまり句点直後の文字から次の句点までの範囲にはならないのが普通である。しかし、上掲の 5 例は文単位で引用してある。どうしてこのようなことができたのだろうか。それは、

- ⑦ 各 KWIC をダブルクリックするとそれを含むファイル全体を別ウィンドウで表示させることができる。

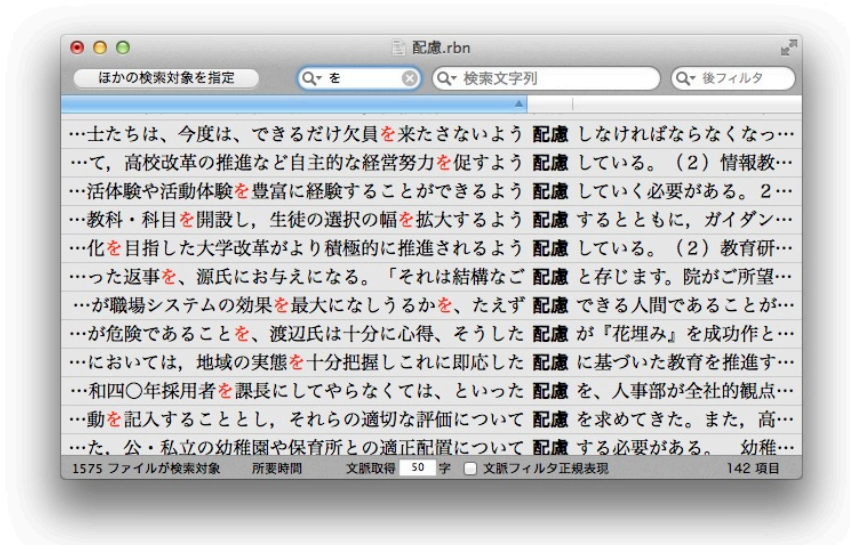
という機能を用いたからである。表から簡単に元のファイルを呼び出すことができるので、そこから文単位で KWIC をコピー／ペーストしたのである。

また、全 235 件の「配慮」の用例からどのように「＜名詞＞を配慮する」の例だけを選び出したのか、という疑問も湧くかもしれない。方法は実は単純で、⑥の機能を使い、前文脈をキーとして表を並べ替えた後、KWIC を上から一つ一つ自分の目で確認していったのである。しかし、その際には同時に次の機能も用いた。

- ⑧ 前文脈・後文脈について条件を指定することで、その（それらの）条件に合う例文だけを表示させることができる（前後文脈フィルタ）。

これはスクリーンショットが分かりやすい（図 4）。

図 4：前文脈に「を」が含まれている KWIC だけを表示する



ウィンドウ上部や左のフィールドに語句を入力すると、その語句を前文脈に含む用例

だけが表示され、合致する部分が赤色で強調される。同様に、ウィンドウ上部右側のフィールドは後文脈の条件を指定することができ、表示される用例をさらに絞り込むことができる。これによって表示されなくなった KWIC は表から消去されたわけではなく単に隠れているだけであり、フィールドの語句を削除すれば再び表示されるので、データ喪失の心配はない。図4では、左のフィールドに「を」を入力することによって、前文脈に「を」を含んだ例のみを表示させたので、表が見やすくなっている。この状態で、筆者は「<名詞>を配慮する」の用例を1つずつ拾っていったわけである。

3 おわりに

本稿で筆者が示したような作業は、もちろんオンラインの各種コーパスとコンコーダンサーを使っても可能である。特に KOTONOHA「現代日本語書き言葉均衡コーパス」は質が高く、これをオンラインで活用できる「少納言」と「中納言」は単語の詳しい用例調査に欠かせない。

しかし日々の授業のために例文を探すとすると、手軽に素早く使えて検索や表示も高速、しかも検索結果を柔軟に活用することができるオフラインのデスクトップ・アプリケーションは依然として存在価値が高い。

今後は授業のための例文収集にとどまらず、学生が書いた作文をコーパスとして誤用の傾向を分析するなど、Simple KWIC Lister の活用範囲をさらに拡大していきたいと考えている¹⁵。

(あきざわ ともたろう／アメリカ・カナダ大学連合日本研究センター常勤講師)

注

- 1 米国 Apple 社製のコンピューターに搭載されている基本ソフトで、従来は Mac OS X と呼称されていたが、バージョン 10.8 (コードネーム Mountain Lion) の登場を機に名称が変更された。
- 2 使用環境が OS X 10.8 以上の場合、インストールされた Simple KWIC Lister を初めて起動する時は、アプリケーションのアイコンを右クリックして「開く」を選択する。アイコンをダブルクリックして起動しようとする、OS のデフォルト設定では起動が許可されない。
- 3 OS X で日本語コーパスの高度な解析が行えるツールとしては、今尾康裕氏(大阪大学)による Casualconc J がある。単純な例文検索はもちろん、他にも多くの高度な機能を備えたソフトウェアである。 (<https://sites.google.com/site/casualconcj/>)

- 4 HDD を SSD に換装した、2007 年発売の MacBook Pro (CPU: Intel Core 2 Duo 2.4GHz、メモリ 6GB、OS X バージョン 10.8.1) を用いて計測した。
- 5 長谷部陽一郎氏 (同志社大学) が、Wikipedia からのコーパス作成の詳しい方法と、そのために役に立つツールを公開されている。 (<http://wp2txt.rubyforge.org>) 筆者もこれを活用させて頂いた。
- 6 注 4 に同じ。
- 7 辞書からの例文への下線は筆者による。
- 8 ②の第 1 段落で使用了のものと同じコーパスを用いた。
- 9 他にも「～に対して配慮する」などの用例が見つかるが、本稿では割愛する。こうした例も含めた、「配慮する」の詳細な用法調査は今後の課題とする。
- 10 山田雄一『稟議と根回し』 (講談社現代新書、1985 年)
- 11 同上。
- 12 吉田寿三郎『高齢化社会』 (講談社現代新書、1981 年)
- 13 井上昌次郎『睡眠の不思議』 (講談社現代新書、1988 年)
- 14 コナン・ドイル (延原謙 訳) 『シャーロック・ホームズの冒険』 (新潮文庫、1953 年)
- 15 読者諸氏にもぜひ Simple KWIC Lister をご活用いただき、akizawa@iucjapan.org 宛にご意見を賜うことができれば幸いである。

参考リンク

- ・ コーパス日本語学のための情報館 (<http://www30.atwiki.jp/corpus-ling/>)
李在鎬氏 (筑波大学) による、コーパスならびに検索・解析ツールの紹介。充実している。
- ・ コーパス言語学入門 (<http://language.sakura.ne.jp/s/corpus.html>)
石川慎一郎氏 (神戸大学) による、英語研究のためのコーパス活用入門。主要コンコーダンサー (Windows 版) の詳細な説明がある。