日本語フレームネットコーパスおよび検索ツール

斎藤博昭（慶應義塾大学）

本稿では，日本語フレームネット（JFN）の分析・アノテーション用コーパスと検索ツールについて述べる(1).

1．JFN コーパス

目下 JFN で使用しているコーパスデータは書き言葉であり，主に新聞記事から成る．新聞記事は，自然言語処理ツールの多くが言語モデルとしてそれを対象としていることからもわかるように，書き言葉としての質が比較的均一であり，大量の言語資源として適している．1 年分の『CD-毎日新聞（データ集)』にはおよそ 100 万文が含まれており，JFN では現在毎日新聞 1992 年から 2002 年までの 11 年分を対象としている．

しかしながら，これまでの「移動」と「伝達」動詞のパイロット分析において明らかになってきたことは，新聞記事コーパスデータのみではジャンルに偏りがあると言うことである．たとえば,「移動」フレームに関与するLU「めぐる」の例文を新聞記事コーパスから抽出しようとしても,「印パ両国が領有を巡って争うカシミール問題」,「県警は暴力団を巡る相談などを電子メールで受け付けている」,「ドイツ軍・オーストラリア軍とロシア軍がここを巡って一メートル刻みの激しい白兵戦を繰りかえした」など,「移動」フレームではなく「主題」フレームに関係する例文が大半を占めてしまう．また,（鈴木 2005）にあるように，評価を伴う伝達動詞「ほめる」,「しかる」,「おこる」の分析では，英語では対応する意味フレームのフレーム要素として定義されていないMessageというフレーム要素が多くの新聞コーパスからの例文で言語化されていることは，新聞というジャンルの特性と無関係ではないと考えられる．従って，目下小説やエッセイも適宜参照しているが，多様な意味分野を分析するのにふさわしいバランスの取れたコーパスとはどのようなものか，今後も雛型作成作業を通して検討していく．

2．JFN kwic ツール

このような大量のコーパスから表現例を探すにはコンピュータ上のツールが不可欠で， JFNでは独自の検索ツールJFN kwicを開発してきた．このツールの特徴は， CaboCha(2)と組み合わせて使うことである．そのため，CaboChaが許す形式のファイルであればどんな形式のものもJFNコーパスに取り込めることになり，基本的に無限にコーパスを拡充できる．また，高速化の実現のために以下のような考慮をしている．すなわち，生コーパスを前もってCaboChaにかけ，形態素解析および係り受け解析をしておく．さらにその解析済みのコーパスにインデックスを付けておく．そのため実際の検索時にはそのインデックスファイルを通して高速探索ができる．図 1 にシステム構成を示す (図中の点線は前もって処理をしておく作業を表す). CaboChaによる形態素解析により，用言の終止形を入力すればすべての活用形の検索が可能になり，また，係り受け解析により後述するように結合価（述語項関係，ある用言に係る名詞句の伴う格助詞の組み合わせ）情報がある程度は抽出できる(3).

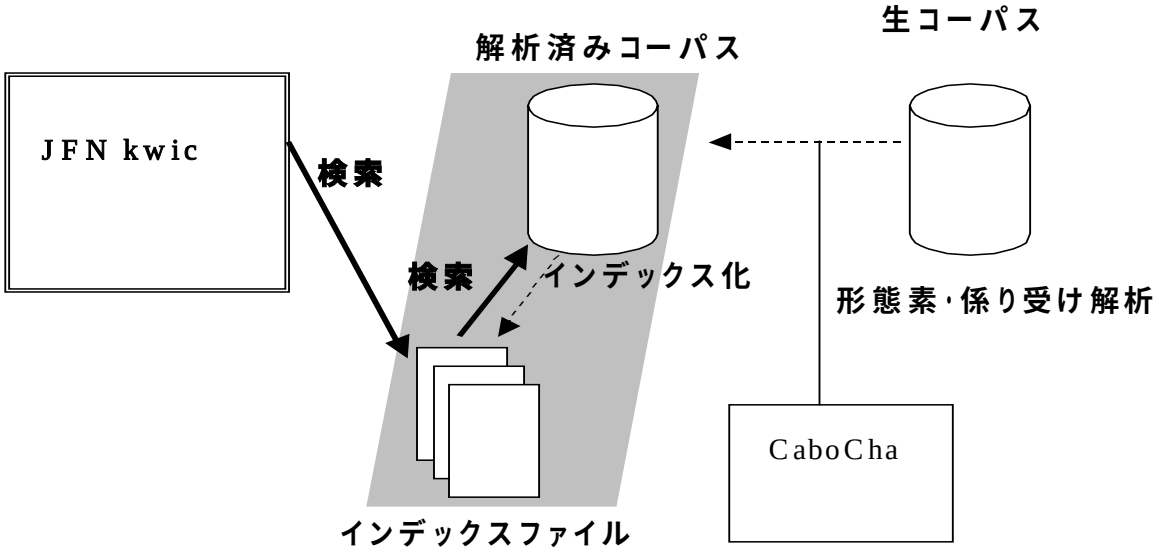

図 1　JFN kwic システムの構成

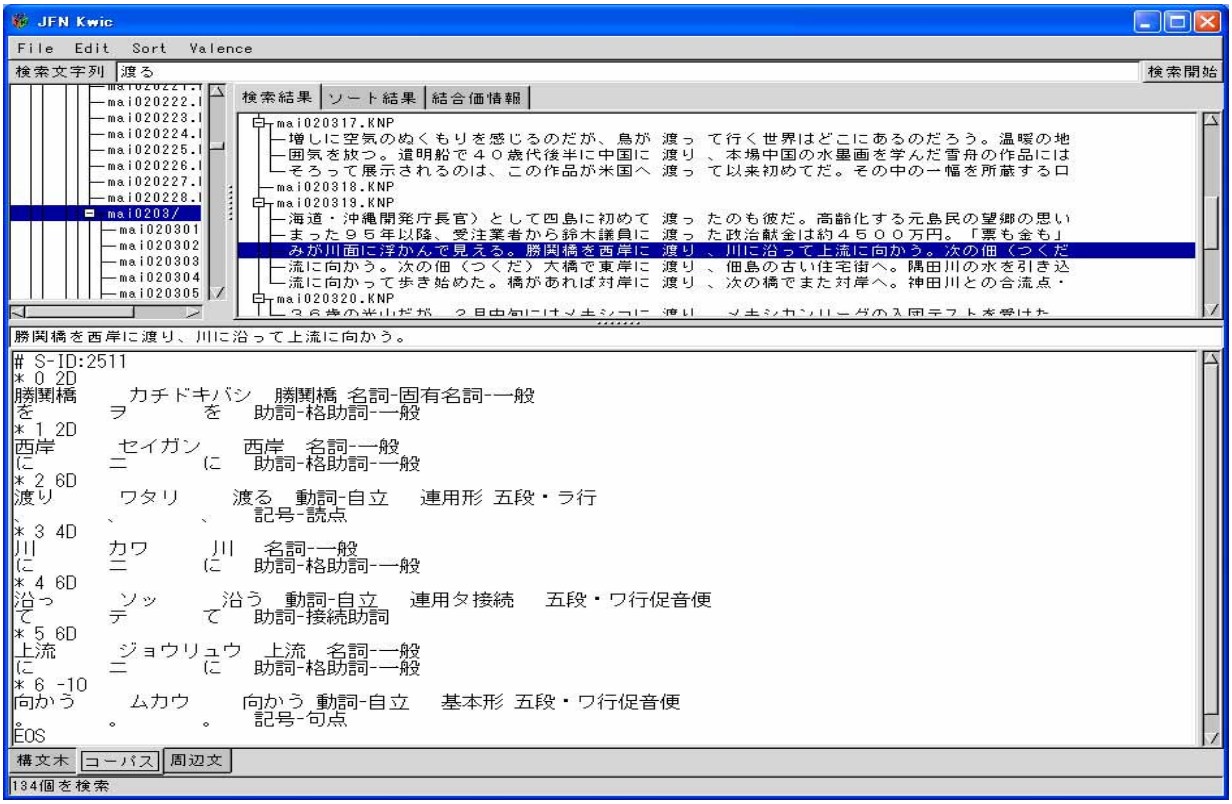

図 2　形態素情報表示画面

ここで，JFN kwic検索ツールの使用例を具体的に見ていく．検索の目的が対象LUの文中および文脈内での使われ方を分析することなので，KWIC (Key Word In Context) 検索が基本となる．検索した文の表示には，文節の係り受けを表示する，形態素解析結果を表示する（図２），その文の周辺にある文とともに表示するといった３つのモードがある．動詞「渡る」を検索した時の検索語彙に係る格要素を結合価情報として抽出表示し

た例を図３と図４に示す．CaboChaによる係り受け解析精度は新聞コーパスの場合およそ９割であり，これらの図にも解析に失敗して誤って抽出された結合価情報やその例文が含まれている．しかしながら，（小原他 2005）にあるように，アノテーション作業の対象例文としては，LUごとにできる限り多様な項関係を持つ例文を集める必要がある．従って，CaboChaの係り受け解析器を利用したこの結合価情報抽出機能は，アノテーション対象例文取捨選択プロセス(4)の前処理としては役に立つ(5)．

謝辞

ツール作成にあたっては，慶應義塾大学大学院理工学研究科の学生であった小杉拓也氏と白川英晃氏の多大な協力があった．深く感謝します．

注

(1) JFN データベースならびにアノテーション支援ツールは，すでに FN のものを移植し日本語化作業を終えているが，これらについては別の機会に報告する．

(2) http://chasen.org/～taku/software/cabocha/

(3) ここで言う「結合価情報」とは，（小原他 2005）の「結合価パターン」と異なりフレーム要素に関する情報は含まず，助詞などの表現形式に関する情報のみを指す．

(4) FN では，このプロセスを，アノテーション作業用サブコーパスを作成するプロセスなので，“subcorporation”と呼んでいる（Fillmore et al. 2003）．

(5) 実際，「移動」領域の「渡る」の結合価情報として，『日本語語彙体系CD-ROM版』には「を+に+渡る」は載っていないが，図４のハイライトされた例文「勝鬨橋を西岸に渡り，…」が示すように毎日新聞記事を検索した結果，このパターンを抽出することができた．

参考文献

Fillmore, Charles J., Petruck, Miriam R.L., Ruppehofer, Josef., and Wright, Abby. 2003. “FrameNet in Action: The Case of Attaching” International Journal of Lexicography Vol.16, No.3, pp.297-332.

小原 京子，石崎 俊，大堀 壽夫，斎藤 博昭，鈴木 亮子，藤井 聖子, 2005.「日本語フレームネット概要」日本認知言語学会論文集第５巻 ＪＣＬＡ 5.

鈴木 亮子, 2005.「評価を伴う伝達動詞：ほめる・しかる・おこるの分析」日本認知言語学会論文集第５巻 ＪＣＬＡ 5.

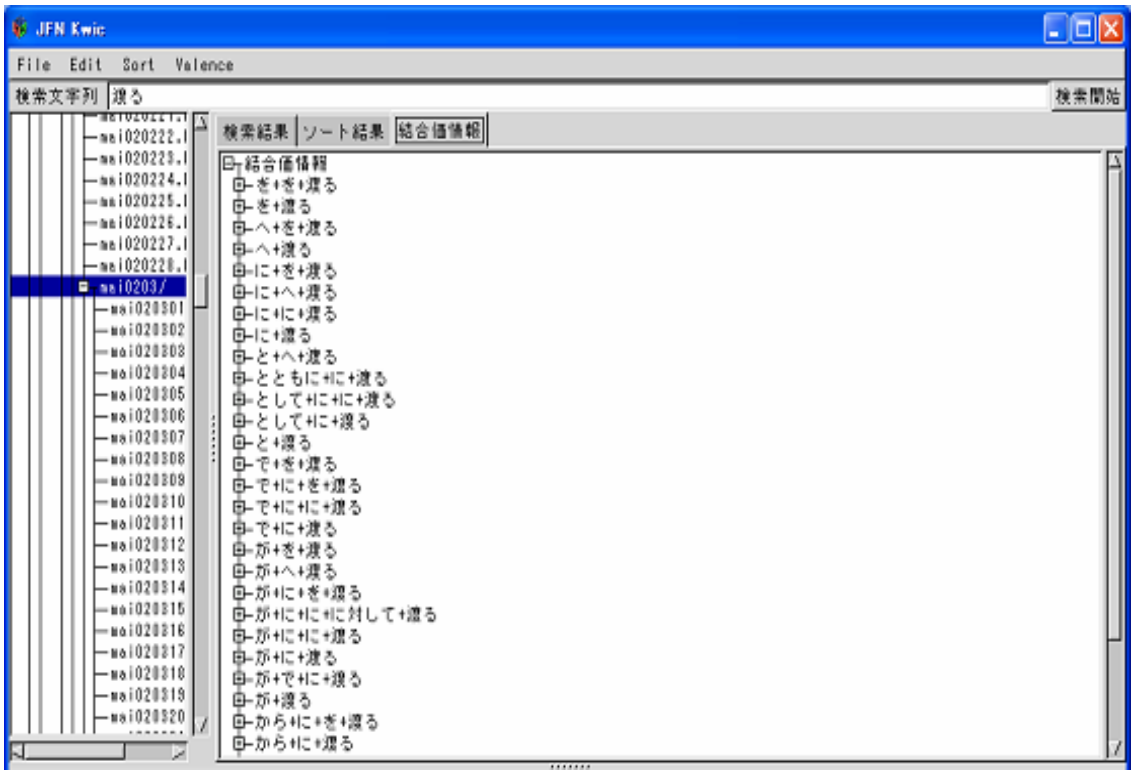

図 3　結合価情報表示画面(ノード閉)