電子化辞書を用いた、文系研究者向き言語分析

荻野 孝野　三吉 秀夫　小林 正博
Takano OGINO　Hideo MIYOSHI　Masahiro KOBAYASHI
(株) 日本電子化辞書研究所
E-mail: {ogino, miyoshi, kobayasi}@edr.co.jp

1. はじめに

EDR (日本電子化辞書研究所略称) では、1986年より9年間の電子化辞書開発を行ない、その後、1995年より保守改良を継続的に行なっている。

開発した電子化辞書は、大学研究機関(国内104件、海外8件)、公的研究機関 (国内6件、海外1件)、民間及び国家プロジェクト (国内36件、海外2件) などにおいて、言語処理関係の研究開発に活用されてきている。この研究内容については、EDR電子化辞書利用シンポジウム[1,2]、ホームページ(http://www.iijnet.or.jp/edr)などでも紹介されている。

辞書別の利用状況は、筆者らが把握している範囲では表1のような内訳である。

表1　辞書別の利用件数

日本語単語辞書	137
英語単語辞書	80
概念辞書	126
日英対訳辞書	74
英日対訳辞書	69
日本語共起辞書	125
英語共起辞書	76
専門用語辞書	69
合計	756

ただ、これらは工学系の研究者に利用されている事例がほとんどである。本発表では、特に文系研究者にも興味が持てるような、電子化辞書の概要および電子化辞書そのものを対象として言語データ分析を、パソコン上でのデモによって紹介する。以上のような観点で、主に日本語データ分析を対象とした日本語関係の共起辞書、EDRコーパスなどを利用した事例を取り上げる。

2. EDR 電子化辞書概要

まず、EDR電子化辞書そのものについて、簡単に構造の概要などを解説する。2.2の各部詳細については、本稿でふれる日本語の分析を中心にした利用を想定し、日本語単語辞書、共起辞書、コーパスなどの範囲にふれる。

2.1 全体構成

EDR電子化辞書は、日本語単語辞書 (25万語)、英語単語辞書 (19万語)、日英対訳辞書 (23万語)、英日対訳辞書 (16万語)、概念体系辞書 (40万概念)、概念記述辞書 (40万概念)、日本語共起辞書 (90万句)、英語共起辞書 (46万句)、専門用語辞書 (情報処理分野、日本語専門用語単語辞書-12万語, 英語専門用語単語辞書-8万語)、日本語コーパス (22万文)、英語コーパス (16万文)、から構成される。

2.2 各辞書の詳細

2.2.1 日本語単語辞書

(1) レコード数：約40万レコード

(2) レコード形式:

レコード番号、見出し情報、文法情報、意味情報、運用・その他情報および管理情報から構成されている。

(3) 出力例：

JWD0186811　保守する[ホシュ・スル]
保守(JLN3,JRN4)　ホシュ　ホ'シュ　JN1;JVE
JA13　JA22　JK01　JK02　JK04　JRN4　108b73
{maintain}　["to maintain the normal condition of a machine"]
{保守する[ホシュ・スル]}　[機械などの正常な状態を保守する]　DATE="94/6/7"

(4) 利用

単に参照の辞書としての利用[3]は、もっとも基礎的な利用であるが、EDR辞書の特色はパソコンなどに搭載されている参照用辞書とは違ったレベルで、連接属性などの文法情報を用いた形態素解析などに利用されている[4,5,6,7]。

2.2.2 日本語共起辞書

日本語共起辞書は、構文、意味の両方の観点から二項の係り受け関係を抽出し、係り受け関係にある文の構成要素のセットと、その関係を記載したものである。

(1) レコード数：約1144万レコード

(2) レコードの形式:

レコード番号、句見出し、共起句構成要素情報、構文情報、意味情報、共起状況情報、および管理情報から構成されている。

(3) 出力例：

共起辞書に含まれる共起関係は、例3.3を参照のこと。

(4) 共起辞書の利用

共起辞書は、格関係でみた用言の分析などのデータ

として、松本氏ら（奈良先端大学）による「格パターン分析に基づく動詞の語彙知識獲得」などに用いられている[11,12]。他にも共起辞書とコーパスを用いた「格フレーム抽出」事例が報告されている[13,14]。

本発表でも、ＥＤＲ辞書の利用事例のひとつとして、共起辞書から抽出したデータで、用言の共起関係パターンを抽出する過程を紹介する。

(5) 日本語動詞共起パターン副辞書

約５千語の主要動詞について、概念見出しごとに動詞の格に関連する共起パターンを、表層格、深層格関係で記述したもので、各パターンごとに例文が付与された約１万４千レコード（１行１レコード）のデータである。

出力例はデモで紹介する。

2.2.3 日本語コーパス

(1) レコード数：約２１万レコード

(2) レコードの形式：

テキスト番号, 出典情報, 形態素情報,構文情報, 意味情報から構成されている。

(3) 出力例：付録１

(4) コーパスの利用

付録１に示すように、コーパスは、1)形態素情報(形態素ごとに品詞および概念識別番号がついている）、2)（ ）で示す構文的な係り受け関係、3)agent,objectと言った概念関係による意味的な係り受け関係、の３種の情報を記載したものであり、構文的な係り受け関係を抽出整理して文法作成の検討に用いた利用例などが報告されている[10]。また、意味的な係り受け関係を用いれば、用言の深層格パターンの検討などにも利用可能である。

2.2.4 概念体系辞書

概念体系辞書は、単語辞書に語義として導入された40万の概念について 、概念の上位下位関係を用いて体系化したものである。

(1) レコード数：約４０万レコード

(2) レコードの形式

概念体系レコードは、レコード番号、上位概念識別子、下位概念識別子、および備考からなる。

(3) 出力例：例3.9に示す。

(4) 概念体系辞書の利用　単語の意味の類似度の検討、多義性の解消[15]、 情報検索の拡張[16]などに用いられている。

3. 辞書データの利用

3.1 連接テーブル作成のための連接状況調査事例

以下の事例は、形態素解析の連接テーブルを想定し、一見、同じ接続関係をとるように見える、副詞相当の「ゆっくり、さっぱり、そっくり」について、次にどんな語がきているかを実際のコーパス事例から抽出し、検討し、接続テーブルを作成するものである。

例3.1　「そっくり」の次にくる語（品詞）の事例

黒い髪と大きな瞳、日本女性【そっくり】だ<語尾>。
色こそ違うが、ゆりかご会の制服【そっくり】だっ<語尾>た。
最近のロボットは、人間【そっくり】で<語尾>あることを必ずしも必要としない。
記者には、民主、共和両党の大会【そっくり】と<語尾>映ったようだ。
像を模写したとき、目、顔、手が【そっくり】な<語尾>のに気づき、それを示すために描いたという。
達、ネバール陸軍の一部が解放軍【そっくり】に<語尾>もなった。
のそでには、ナチスのかぎ十字に【そっくり】の<語尾>マークがついている。
畳んだ姿は、まったく、枯れ葉に【そっくり】です<助動詞>。
ルバーから聞き出して、わざわざ【そっくり】まね<動詞>たのだ。
きみの経験が【そっくり】そのまま<副詞>　他の人にあてはまるとは限らない。

次に例3.1と同様、「さっぱり、すっかり」についても後続品詞を得る。自立語については品詞、付属語については、付属語に相当する文字列そのままを用い、下記の表を作成する。

例3.2「そっくり,さっぱり,すっかり」の接続表

後続語	そっくり	さっぱり	すっかり
名詞	○	○	○
連体詞	○	○	○
動詞	○	○	○
形容詞	×	○	○
形容動詞	×	○	○
副詞	○	×	×
する	×	○	×
だ	○	×	×
な	○	×	×
の	○	×	×
に	○	×	×
と	○	○	×

3.2 動詞の結合パターンの分析

以下の事例は、動詞の結合価パターン[17]検討のために、EDRコーパスを用いて、動詞にかかる格関係に関するデータを抽出するものである。

例3.3 共起辞書における単語「作成」部分のレコード

JCC0973436
構文情報: 法案(名詞)[108625] =>[が]=>
作成(動詞)[3ce781]

意味情報: 作成(3ce781) =>[object]=> 法案(108625)
例文情報: 006000009dce-7-4/"<法案>が(作成)された"

JCC0991049
構文情報: 本部(名詞)[109016] =>[が]=>
作成(動詞)[0f9516]
意味情報: 作成(0f9516) =>[agent]=> 本部(109016)
例文情報: 00050004d277-26-23/"<本部>が(作成)する"

例 3.4 共起辞書の例文情報で指示するコーパス例文をつけたもの

作成 <= が(法案[名詞])
#だがこの法案が作成された５９年末から国会に提出された６０年３月にかけて、この法改正の直接の責任者だった。
作成 <= が(本部[名詞])
#同社の説明によると、時刻表の数字は、東京圏運行本部が作成するグラフ化された運行ダイヤグラムがもとになる。

例 3.5 例 3.4 のデータの係り受け関係部分のみを取り出したもの

作成 <= が(法案[名詞])
作成 <= が(本部[名詞])
作成 <= が(利用者[名詞]) φ(直接[副詞])
作成 <= が(利用者[名詞])
作成 <= が(利用者[名詞])
作成 <= から(システム設計[名詞])
の(プログラム[名詞])
作成 <= から(データ[名詞]) は(これ[名詞])
を(フォーマット[名詞])

以上で抽出したデータを元に以下のような格関係が整理される。ここでは、活用語尾まで抽出していないので、形上、[名詞]部分にくる単語の意味によって、受身を識別することになるが、それ以前に、活用語尾までを抽出するようなデータ抽出とすればより適切な処理となる。

例 3.6

作成 <= が(<人>[名詞])
を(<物>[名詞]) から(<物>[名詞])
作成 <= が(<物>[名詞])

3.3 共起辞書と概念体系を用いた概念関係表現の抽出

EDR 概念辞書は、概念体系辞書と概念記述辞書からなる。概念記述辞書とは、文の構成要素間の意味的な係り受け関係を「agent, object, source,」といった関係で記述したもので、いわゆる用言の深層格相当の情報を記載したものである。

ここでは、EDR コーパスに記載された例「野菜を食べる」の意味情報から、体系の中間ノードレベルの概念関係表現に抽象化していく過程を示す。

3.3.1 共起辞書に出現する「野菜を食べる」

例 3.7

JCC1067445
構文情報: 野菜(名詞)[10c06f] =>[を]=>
食べ(動詞)[3bc6f0]
意味情報: 食べ(3bc6f0) =>[object]=> 野菜(10c06f)
例文情報:
00050007e0be-24-21/"<野菜>を(食べ)ていった"

3.3.2 上位の概念関係へのシフト

次は、概念体系によって、「野菜」の上位「飲食物」と「食べる」の関係を得る。

例 3.8 --- 基本語辞書 (完全一致検索)-[野菜] -------

JWD0202844 野菜[ヤサイ] 野菜(JLN1,JRN1)
ヤサイ ヤサイ JN1 10c06f {""} ["a plant grown for food and eaten with the main dish, called a vegetable"]
{野菜[ヤサイ]} ["畑などに作って収穫し,副食物にする植物"] DATE="89/2/17"

例 3.9 概念体系における「野菜」の階層関係

00: 3aa966 概念
01: 3d017c ものごと
02: 444d86 もの
03: 30f6ae 具体物
04: 30f6af 生命体
05: 30f6ca 植物
06: 3aa91d 役割で捉えた植物
07: 30f6d2 食用植物
08: 30f6cf 野菜
09: 10c06f "畑などに作って収穫し,副食物にする植物" [野菜]
04: 4444c4 静物
05: 3aa92f 機能で捉えた具体物
06: 3f9639 飲食物
07: 44457a 食品
08: 44457b 天然食品
09: 30f6cf 野菜
10: 10c06f "畑などに作って収穫し,副食物にする植物" [野菜]

上記の体系によって、「野菜」が「飲食物」および「植物」の下位に相当することがわかる。「飲食物」の下位にあるその他の事例についても、「食べる」との関係が object であることが確認されれば、「食べる=>[object]=> 飲食物」の関係を導くことができる。ここでは一事例を示すにとどめるが、3.2 で取り上げた「表層格パターンの抽出」同様、用言の深層格パターンの整理が可能である。これらを用いて、「とうもろこしは馬も食べる」のように、係助詞の表現のため格関係が不明になったものの意味関係を「食べる」の深層格パターンから格関係を

類推することも可能となる。上記の例では、「とうもろこし」も「馬」も意味的には「食べる」の object になることが可能で、このレベルではどちらも「を格」をとることが可能である。さらに、共起辞書や概念体系によって、「食べる」の agent になるケースが「動物」や「人間」であることが確認され、「とうもろこしは（object）馬も（agent）食べる」という概念関係を得ることができる。

4．まとめ

以上のような内容で、電子化辞書の概略と言語データとしての利用事例をパソコン上で紹介する。

付録1：日本語コーパスの出力例

<テキスト番号> 000500017459
<出典情報> .....
<文> 会場は熱気に包まれ、集会後、周辺路上でのデモ行進に移った。

<構成要素情報>

<構成要素番号>	<表記>	<かな表記>	<品詞>	<概念選択>
1	会場	カイジョウ	名詞	3c0841
2	は	ハ	助詞	2621d5
3	熱気	ネッキ	名詞	102ab4
4	に	ニ	助詞	2621d5

<形態素情報> /1:会場/2:は/3:熱気/4:に/5:包/6:ま/7:れ/8:、
/9:集会/10:後/11:、/12:周辺/13:路上/14:で/15:の/16:デモ/17:行進/18:に/19:移/20:っ/21:た/22:。/{I#1//16:デモ/17:行進//}

<構文情報>

```
(S (t (M (S (t (M (S (t (W 1 "会場"))
                         (W 2 "は"))
                    (t (M (S (t (W 3 "熱気"))
                             (W 4 "に"))
                        (t (S (t (W 5 "包"))
                              (W 6 "ま")
                              (W 7 "れ")))))))
```

<意味情報>

```
[  [main 18:移:0e5eca]
   [time [  [main 10:後:3d0476]
            [modifier 9:集会:3cec82]]]
   [goal [  [main I#1:デモ行進:"＝Z　何かに反対して
                行う路上での行進"]
            [place   [  [main 13:路上:10ebf5]
                        [modifier 12:周辺:3cf780]]]]]
```

5．参考文献

[1] 日本電子化辞書研究所；ＥＤＲ電子化辞書利用シンポジウム論文集(1995)

[2] 日本電子化辞書研究所；ＥＤＲ電子化辞書利用シンポジウム論文集(1997)

[3] 増山顕成、塩津誠；英和・和英辞書ソフトウェア「電辞海」について;ＥＤＲ電子化辞書利用シンポジウム論文集(1997)

[4] 黒橋 禎夫、長尾 眞；日本語形態素解析システム JUMAN Version3.4;(1997.11) http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/

[5] 植木正裕、徳永健伸、田中穂積；EDR 辞書を用いて日本語文の形態素解析と統語解析を行なうシステム;ＥＤＲ電子化辞書利用シンポジウム論文集

[6] 相場徹、奥村学;構文・意味解析と統合した形態素解析に関する研究;ＥＤＲ電子化辞書利用シンポジウム論文集(1995)

[7] 永田昌明;EDR コーパスを用いた確率的日本語形態素解析;ＥＤＲ電子化辞書利用シンポジウム論文集(1995)

[8] 吉村裕美子;EDR 対訳辞書を用いた英日機械翻訳用辞書の大語彙化;ＥＤＲ電子化辞書利用シンポジウム論文集(1997)

[9] 藤尾正和,松本裕治;EDR 括弧付きコーパスを利用した、統計的日本語係り受け解析;ＥＤＲ電子化辞書利用シンポジウム論文集(1997)

[10] 白井清昭、徳永健伸、田中穂積;EDR コーパスからの確率文脈自由文法の自動抽出に関する研究;ＥＤＲ電子化辞書利用シンポジウム論文集(1995)

[11] 大石 亨, 松本 裕治;パターン分析に基づく動詞の語彙知識獲得;情報処理学会論文誌, 第36巻, 第11号, pp.2597-2610, (1995.11)

[12] Akira Oishi and Yuji Matsumoto;AMethod for Deep Case Acquisition Based on Surface Case Pattern Analysis,Proceedings of the 3rd Natural Language Processing Pacific Rim Symposium, pp.678-684 (December 1995)

[13] Tsunenori Mine, Masaru Higashi and Makoto Amamiya;Case Frame Acquisition and Verb Sense Disambiguation on a Large Scale Electronic Dictionary,Proc. of NLPRS

[14] 橋本順子、峯 恒憲、雨宮 真人;和語動詞の格フレームを利用したサ変動詞の格フレーム獲得;情報処理学会全国大会講演論文集(1997)

[15] 内山将夫、板橋秀一；シソーラス上での共起頻度を利用した動詞の多義性解消；情報処理学会研究報告 96-NL-116, Vol.96, No.114, 1996.

[16] 太田千晶 奥村学;EDR 電子化辞書を用いたクエリー拡張による検索支援;言語処理¥学会 第3回年次大会発表論文集(1997)

[17] 荻野孝野;日本語処理における結合価文法とその問題点;朝倉日本語新講座3 文法と意味1;朝倉書店(1983)