情報メディア 34－7
（1998. 11. 27）

ＥＤＲ電子化辞書について

----言語資源の観点で------

荻野 孝野
日本電子化辞書研究所
ogino@edr.co.jp

昭和５０年代初期において、冊子体の辞書が電子化された。その後、冊子体辞書の構造化時代を経て、自然言語処理をさらに進展させるために辞書を開発しようという動きの中で、昭和６１年ＥＤＲ電子化辞書プロジェクトがスタートした。現在、これらの辞書は、言語処理に携わる研究者によって、各システムにあわせた言語データとしてさらに加工して利用される言語資源の一つとなっている。本稿では、ＥＤＲ電子化辞書の特徴を「概念」をつなぎ手とする相互関係の辞書ととらえ、概念に着目したＥＤＲ辞書の概要とその利用事例について、言語資源の活用という見方で述べる。また、インターネット普及時代において、電子化辞書をグロバールな利用とするためにかかわっている「オントロジーの標準化」のためのシソーラスの照合などについても紹介したい。

EDR Electronic Dictionary

---- From a viewpoint of language resources —

Takano Ogino
Japan Electronic Dictionary Research Institute, Ltd. (EDR)
ogino@edr.co.jp

In the 1970s several written dictionaries were stored into electronic media as machine readable dictionaries. Since then there has been a movement to develop machine tractable dictionaries that were aimed at achieving an advanced natural language processing technology. Then the national project to develop EDR Electronic Dictionary began in 1986.

Currently EDR Electronic Dictionary is used as an language resource for language analyses by natural language processing researchers. One of the main features of EDR Electronic Dictionary is that its subdictionaries are interconnected via concepts. This paper describes the usage examples of EDR Electronic Dictionary from a viewpoint of language resources. The activity of comparing different thesauri is also described from the viewpoint of glovalization of Electronic Dictionaries.

1. はじめに

９年間の開発およびその後の保守を含めて、ＥＤＲ電子化辞書は、大量の語彙を備えた、構造化された電子化辞書として、多くの言語処理関係の研究者や開発者に利用され、言語処理分野の研究開発の進展を、多少なりとも言語データの側面から支援しているかと思われる。

ここでは、「辞書の言語資源としての活用」という観点から、「電子化辞書がどういう構造を持っているか」「電子化辞書が言語処理のどんな研究に利用されているか」「どういうデータを生み出すか」さらに「電子化辞書をグローバルな利用に広げるための工夫」などについて報告したい。

2. 冊子体辞書から電子化辞書への変遷

筆者が初めて電子化辞書の仕事に携わったのは、昭和５２年、電子技術総合研究所推論研究室において電子化された「新明解国語辞典（三省堂）」の電子化データの誤り修正と構造化データの作成であった。これらについては、文献1,2,3に経過が報告されている。「新明解国語辞典のデータベース化」は、冊子体辞書をそのまま入力して、入力後に中で使われている見出し区切りや語義番号などから構造化できる記号上の手がかりを見つけて構造化データベースの作成を試みたものである。

その後、京都大学ではコンサイス英和辞典を対象にデータベース化する研究が行なわれた（文献4）。これはあらかじめ、見出しファイル、語義ファイルといった構造化されたファイルを設計し、冊子体辞書を入力する時点で構造化ファイルに変換するための手がかりとなる区切り符合を原文とともに入力しておくものであった。

その後、昭和５９年、ＩＣＯＴ内部において電子化辞書検討委員会が発足し、ここでは「冊子体辞書から電子化辞書を」ではなく、最初から「構造化された、電子化辞書のための辞書」を作ろうということで、「電子化辞書の基本的構造の設計」の議論が行なわれた(文献5)。これらの活動は現在の電子化辞書開発へと広がり、ＥＤＲ電子化辞書の土台となったものである。

以上のように、昭和５０年代初めより、言語処理のための辞書開発の重要性が、機械翻訳など自然言語処理開発サイドから指摘され、電子化辞書そのものの開発が進められてきた。辞書開発は、

1)冊子体辞書を入力した後で、既存の符合をてがかりに構造化

2)冊子体辞書に「構造化のための符合」を付加して入力後、構造化

3)最初から構造化した形式で電子化辞書を作成

というような開発段階を経て今日に至った。

もちろん、家庭でも手軽に入手でき、広く普及している冊子体辞書と共存する「CD−ROM版電子辞書」や「携帯版電卓型電子辞書」は、あくまでも「検索表示し見るための辞書」であり、本稿で述べる「電子化辞書」は、利用目的を異にする「目的にあわせ加工して利用する」ための辞書である。

3. ＥＤＲ電子化辞書の全体構成

上記のような辞書開発の流れの中で、構造化された辞書として、言語処理利用を前提として開発された「ＥＤＲ電子化辞書」の全体構成を図１に示す。

本稿では、図１の中で★印のついた辞書を中心にとりあげる。

【日本語単語辞書】

レコード番号、見出し情報、文法情報、意味情報、運用・頻度情報および管理情報から構成されている。通常の冊子体辞書や表示用電子辞書と異なる部分は、形態素解析用の形態素接続テーブルの縦軸、横軸で連接情報を提供する部分などである。

【概念体系辞書】

単語辞書に語義として導入された４０万の概念について、約６０００の分類枠におさめ、これらの分類枠を概念の包含関係に基づいて、概念体系として構造化したものである。

【概念記述辞書】
概念記述辞書は文中で係り受け関係にある概念間（二項）の意味的関係（動作主、道具、場所、など）を整理したものである。
【日本語共起辞書】
文中で係り受け関係にある単語どうしについて、つなぎとなる助詞などを「共起関係子」として介在させた二項関係で、文中の表層的な係り受け関係を構文情報として、深層的な係り受け関係を意味情報として記載したものである（末尾ページの資料参照）。
【日本語コーパス】
大量の用例に、形態素情報、構文情報、意味情報などを付与した総合的な言語データである。

図1　ＥＤＲ電子化辞書全体構成

単語辞書	日本語単語辞書★ 英語単語辞書	２５万語 １９万語
対訳辞書	日英対訳辞書 英日対訳辞書	２３万語 １６万語
概念辞書	概念体系辞書★ 概念記述辞書★	４０万概念 ５１万組
共起辞書	日本語共起辞書★ 英語共起辞書	９０万句 ４６万句
ＥＤＲコーパス	日本語コーパス★ 英語コーパス	２２万文 １６万文
専門用語辞書 （情報処理）	日本語専門用語単語辞書（情報処理） 英語専門用語単語辞書（情報処理） その他（概念体系、対訳、共起の各辞書を含む）	１２万語 ８万語

4. ＥＤＲ電子化辞書の特徴

「ＥＤＲ電子化辞書」の代表的な特徴としては、以下のような点があげられる。

1)「単語見出し」あるいは「概念ID」を共有して、辞書は相互に関係する。

2)単語辞書を用いた単なる辞書引きだけというのではなく、「概念体系辞書、概念記述辞書、共起辞書、タグつきコーパス」などと連動し、「構文レベル、意味レベル」といったより進んだ言語　処理の利用も想定した辞書構成である。

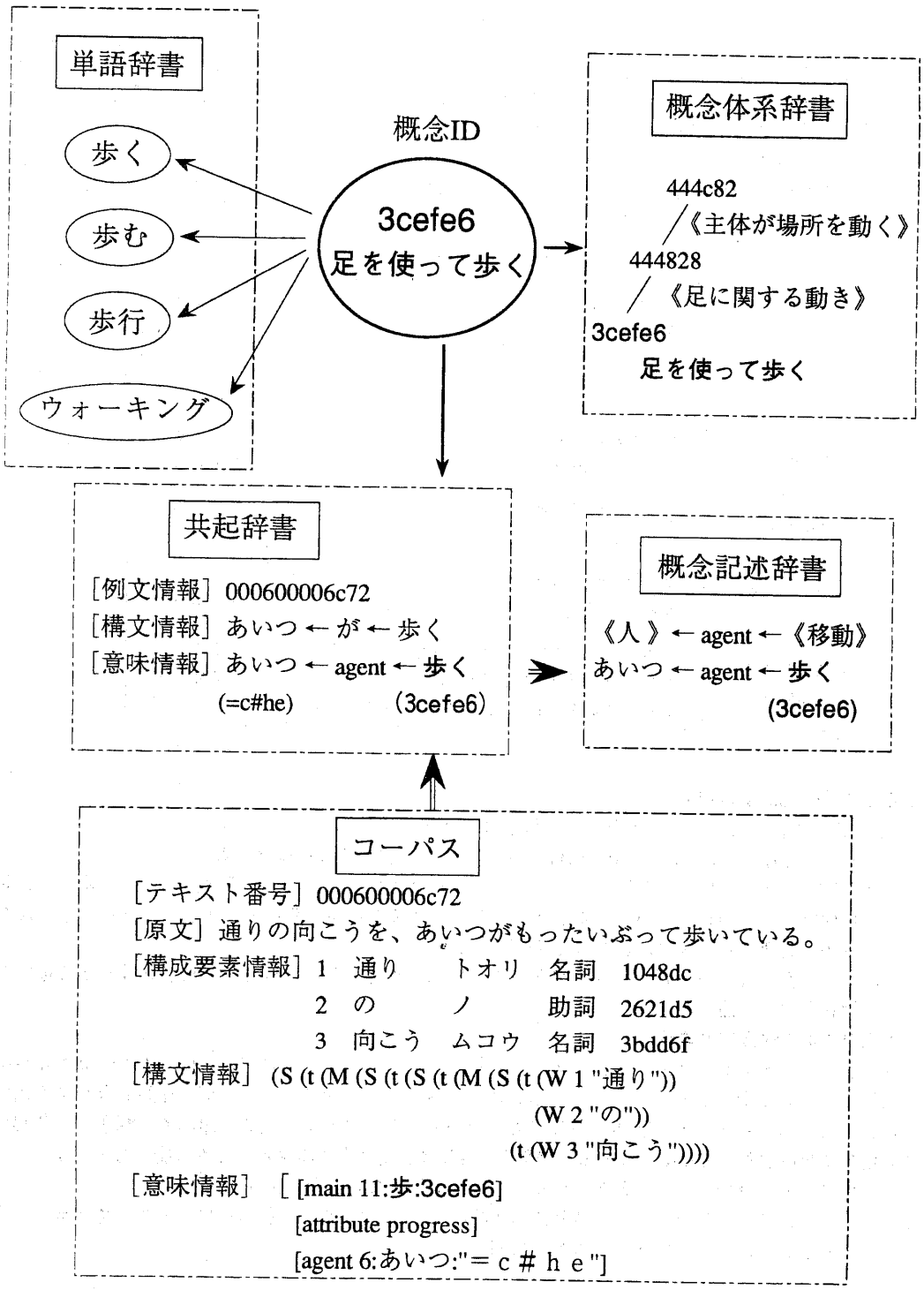
これらの特徴を「データの効率的蓄積からみた辞書の構造」という観点で、冊子体辞書との違いなどにも触れながら、述べる。

・概念IDによる辞書間の連結

「概念ＩＤ」を介して関連する辞書データの様子を図２に示す。この図から概念IDでつながる辞書の関係をみることができる。単語辞書で単語に記載された「概念ID」は、「概念体系辞書」にも用いられ、また「コーパス」の形態素の意味にもつながる。このように、同じ言語内で、概念は、各辞書間で相互に関連する構造となっている。異なる言語間で概念IDを共有することも可能である（ただし、異言語間については徹底した統合まで行っているとは言えない部分も含む）。

また、より具体的に、ＥＤＲで開発されたUNIX用辞書検索ツール（文献11）によって、共起辞書内容の一部を例示しておく（末尾ページの資料参照）。

図２　　概念ＩＤをつなぎ手とした各辞書の関係



単語辞書のいわゆる語義に相当する概念見出しに付与された「概念ID」は、概念体系辞書によって上位概念が何であるか、概念階層の中での位置として認識することができる。

また、コーパスを構成する文は、単語辞書を使った形態素解析によって、辞書の単位に分割され、該当する辞書の品詞が付与され、構文解析ツール、概念解析ツールによって構文上の係り受け関係、概念の選択、概念上の係り受け関係が付与され、単語辞書に出現する単語や概念レベルの出現頻度を把握したり、表層レベル、深層レベルの係り受け関係を抽出できる。

コーパスから二項関係で構文情報や意味情報を抽出したものが「共起辞書」であり、意味情報のみに着目したものが「概念記述辞書」である。これもすべて、単語辞書から発生した概念IDと連結した形で構成されている。

・単語辞書内の概念共有化

図2でわかるように、「3cefe6 足を使って歩く」の概念は、「歩く、歩行、歩む」などの単語の概念にも使われている。このようにＥＤＲ電子化辞書では、異なる単語でも同じ概念IDを概念の説明に使うことができる。冊子体の辞書では、このあたりの作業は徹底化できない部分であるが、電子データでは相互参照して統合可能な部分である。また、同じ概念を複数の単語が共有するという「概念共有化」は、データ蓄積の経済性をねらったものであるが、これは、現在の言語処理のレベルで「概念の判別に必要な程度の概念区別」という粗さも含んだ上での、言語の平易化とも言える。

5. ＥＤＲ電子化辞書の利用研究

現在、ＥＤＲ電子化辞書は、大学や民間の研究所など辞書数で８３３件の利用契約が取り交わされて利用されている。辞書別内訳は表１に示す通りである。

表１　辞書別の利用（1998.10）

	日本語単語	英語単語	概念	日英対訳	英日対訳	日本語共起	英語共起	専門語	合計
利　用	155	85	142	80	74	139	81	77	833
機関数	18.6	10.2	17.0	9.6	9.0	16.7	9.7	9.2	100%

表１から「日本語単語辞書、概念辞書、日本語共起 辞書」などの利用が多いことがわかる。

また、これらの利用状況を、年度（1996～1998)、学会(人工知能学会, 情報処理学会、言語処理学会、電気関係学会)で限定して、ＥＤＲ電子化辞書利用の研究発表１４７件でみた利用の内訳を表２に示す。

表２　学会発表論文の辞書別利用状況

EDRコーパス	97	概念辞書	37	英単語辞書	9	日英対訳辞書	2
日本語単語辞書	48	日本語共起辞書	28	英日対訳辞書	3	専門用語辞書	2

「EDRコーパス、日本語単語辞書、概念辞書、日本語共起辞書」の利用が顕著である。特に「ＥＤＲコーパス」については、形態素処理結果のチェックコーパスとして、あるいは、N-gram形態素解析で品詞や文字列連鎖を統計的にとるデータとして利用された事例（文献9)などが報告されている。また、形態素情報と意味情報を用いた「機械翻訳のための訳し分け」などのデータ抽出（文献12）や、「括弧つき係り受けの構文情報から文法規則データの抽出」（文献10）などにも利用されている。更に「格パターン分析に基づく動詞の語彙知識獲得」（文献16）など、言語処理のための言語データの抽象化やパターン化のために幅広く活用されている。

「日本語単語辞書」は、形態素ごとの連接情報や連接テーブルを付帯していることもあって、形態素解析用辞書として活用されている事例が多い。京都大学長尾研究室「JUMAN」（文献6）、NTT「すもも」（文献7）、東京工業大学田中・徳永研究室「形態素・構文解析用ツール・MSLRパーザ」（文献8）など形態素解析処理システムに有効に利用されている。

「概念辞書」の一つである「概念体系辞書」は、機械翻訳の訳語選択、仮名漢字変換の同音異語の表記選択などに利用されている。今後の利用として、情報検索の検索用語の絞り込みや拡大などへの利用もあげられる（文献15）。

開発初期の目的として、ＥＤＲ電子化辞書は、概念を共有した「多言語翻訳」のための辞書を想定したものであるが、上記の利用概要からわかるように「加工された言語資源」として活用され、新たな言語処理システムに合わせた「言語データ」として発展的に拡張利用されていることがわかる。

6. ＥＤＲ電子化辞書を基に新しい言語データの作成

現在、ＥＤＲでは、日本語共起辞書から「結合価データ」の作成を行なっている。これは、ＥＤＲ共起辞書に含まれる用言（ここでは動詞および形容動詞、9,400語、14,400語義）を対象とし、表記、IDともに一致の用言にかかる共起関係を統合し、用言にかかる「体言＋格関係子」のセットとして作成するものである。

これらの結合価データは、「きしゃがきしゃできしゃする」のような仮名漢字変換システムの同音異語の表記選択をはじめ、機械翻訳の訳語選択（例１）などに有効活用が想定される。

例１．切る

1)<具体物>を<刃物や鋭利な物>で切る。{cut}
2)通信先と通話していた<電話>を終了する。{hang up}
3)稼働状態の<機械>などを止める。{switch ～ off}
:

ここで作成している結合価データは、例２に示すように、従来の結合価表示だけにとどまらず、概念体系辞書を用いて、体言部分の概念を体系上の中間ノードへ位置付けることによって概念レベルを抽象化して表示したり、文番号によるコーパスリンクによって例文表示も行なう。作業は、

1)共起辞書から結合価データの作成
2)自動作成された結合価データの人手チェック
3)結合価データの体言部分の概念体系上への位置づけ
4)結合価パターンへの抽象化

の手順で進めている。

2)の作業は、格の変換の記載や元データであるコーパスの用言部分の概念選択の誤りの指摘などを行なうものである。後者の誤り指摘はコーパスや共起辞書の改良にもつながる。

例２　##### ア・ク[開く] 開/1e8483{"(扉が)開く"}

が	を	に	へ	から	より	まで	で	と	例文
が(扉(名詞)[100928])									重そうな扉が、きしみ音もなく開いた。
が(円窓(名詞)[3c0dbd])									机の真ん中には、カセットテープレコーダーがあり、前には、円窓が開いている。

:

---------- 体言部分の概念体系上における位置 ----------

03:　30f6ae　具体物　:　:
08:　30f6f9　建具
09:　0ffdc2　"窓,家具などについた,開閉の建具"★
10:　100928　戸　[扉<が>]

7. ＥＤＲ言語データのグローバル利用に向けた活動

ヨーロッパやアメリカでは、電子化された言語データを言語資源として共有化していこうという動きが活発化している。１９９８年５月には、ELRA(European Language Resources Association: 言語資源の普及流通をめざし、EC加盟諸国で作られた組織）主催でFirst International Conference on Language Resources and Evaluation (LREC'98)が開催された。ここでは５００名近くの参加者を得て、言語資源の共有化や標準化に関する一般発表や議論が行なわれた。このように世界的に言語資源を活用していこうという共通認識ができつつある。

7.1 ＥＤＲ概念体系とWordNetとの照合

ＥＤＲでは、1996年3月に、米国ANSI関係の「オントロジー標準化」ad-hoc委員会とのかかわりを持ち、ＥＤＲ概念体系とWordNetとの照合を試みている(文献13)。これは、この委員会での「既存のシソーラスから核となるいくつかのシソーラスを選び、概念の類似性をみながら照合させ、標準的なOntologyを作ろう」という了解に基づく試みである。例３に示すように、「概念の示す範囲が一致するもの、ずれるもの」などがあるが、相互の関係をレベル分けしながら、関係をつけられるところまでは、それぞれのシソーラスを関連づけていこうというのが基本姿勢で取り組んでいる。

ヨーロッパでも、EC-funded projectとして、英語、ドイツ語、スペイン語、イタリア語、オランダ語を対象とした、Euro WordNetの構築が進行している(文献17)。

例３　EDRとWordNetの照合

EDR	WORDNET
1-2　動物　30f6bf 1-2-1　種で捉えた動物　30f6c1 1-2-3　生息場所で捉えた動物　444858 1-2-2　役割で捉えた動物　3aa91a	{animal, animate being, beast, brute, creature, fauna} =30f6bf =>{vermin }<3aa91a =>{work animal }<3aa91a =>{domestic animal }<3aa91a =>{pet}<3aa91a =>{marine animal, sea animal }<444858

7.2 電子化辞書記述フォーマットの標準化提案に添った辞書の開発

ＥＤＲでは現在新聞雑誌などに登場する語でＥＤＲ辞書にない語について、追加語辞書の作成を試みている。これらは、ＩＰＡ創造的育成事業の中で取り組まれた「日本語情報海外発信促進のための言語知識コンテンツ蓄積・流通ソフトの開発」の中で提案された機械翻訳用の辞書データの流通、相互利用のためのUPF(Universal PlatForm)の枠組での辞書記述をめざす予定である。現在、部分的に未完成であるが、約１８０００語の日本語部分の記述が蓄積されつつある。

8. 最後に

以上、本稿では、ＥＤＲ電子化辞書の概要と、言語資源としての活用などについて概観してきた。今後はインターネットの大規模な普及に伴い、情報の共有化を想定した言語データの活用がますます重要なポイントになってくると思われる。本稿でもふれたように、辞書開発および保守は、そういう要求に柔軟に対応していく姿勢が必要であろうと認識している。

また、EDR辞書については、更なる新しい言語資源としての拡張と、現在の不備の保守改良も含め、枯渇させない、活きている辞書として存続させていくことが、開発に携わったものの希望でもある。

資料　　辞書内容の参考表示

以下は、ＥＤＲで開発されたUNIX用に辞書検索ツールを用いて表示したものである。本ツールは、「ＩＰＡ創造的ソフトウェア育成プロジェクト」の一環として作成され、公開されている（文献11）。単語辞書検索GetWORD,概念体系辞書検索GetCPC,共起辞書検索GetCOOC, コーパス検索GetCORPUSなどがあるが、スペースの都合上、「日本語共起辞書」の検索例を上げておく。なお、検索コマンドの後ろについた{ }は検索コマンドの内容を付記したもので、コマンドそのものあるいは結果の表示の一部ではない。

【共起辞書の検索例】：検索ツール GetCOOC

```
>>> Input command  (Current Record = 0)
fu  3cefe6                                  {受け側概念ID 3cefe6　で共起辞書を検索}
>>> Input command  (Current Record = 460)   {460レコードあった}
sr が                                       {共起関係子「が」で二次検索}
>>> Input command  (Current Record = 28)
JCC0018257    構文情報:   あいつ(名詞)[=c#he]   =>[ が ]=>   歩(動詞)[3cefe6]
              意味情報:   歩(3cefe6)   =>[ agent ]=>  あいつ(=c#he)
              例文情報:   000600006c72-19-8/"<あいつ>が…(歩)いている"
JCC0020883    構文情報:   赤ちゃん(名詞)[3d0466]  =>[ が ]=>   歩(動詞)[3cefe6]
              意味情報:   歩(3cefe6)   =>[ agent ]=>  赤ちゃん(3d0466)
              例文情報:   00070000c8bc-9-0/"<赤ちゃん>が…(歩)く"
```

文献

1）横山晶一：国語辞典データベース化の準備,電子技術総合研究所彙報,第４１巻１１号,(1977)

2）荻野孝野：国語辞典ファイル化作業,計量計画研究所報告'８１,(1982)

3）横山晶一,荻野孝野：国語辞典磁気テープのドキュメント,電子技術総合研究所彙報,第４８巻８号,(1984)

4）新コンサイス英和辞書「利用手引書」,昭和55,56年度科学研究費補助金試験研究(1)研究成果報告書,｀言語辞書活用のための計算機プログラムシステムの開発と言語辞書の解析,(1979)

5）H.Miyoshi, Y.Tanaka, T.Yokoi, T.shiwata, H.Tanaka, S.Amano, H.Uchida, T.Ogino, 'Basic Specification of the Machine-Readable Dictionary' Icot Technical Report TR-10,(1986)

6）松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真:日本語形態素解析システムJUMAN 使用説明書 version 2.0, NAIST Technical Report, NAIST-IS-TR94025, (1994) , http://www-nagao.kuee.kyoto-u.ac.jp/

7）鷲坂光一, 山崎憲一,廣津登志夫,尾内理紀夫,情報検索のための高速日本語形態素解析システム「すもも」,情報処理学会第54回全国大会,2-59～60,(1997) , http://www.brl.ntt.co.jp/sumomo/

8）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,(1995) , http://tanaka-www.cs.titech.ac.jp/pub/mslr/index.html

9）森信介,長尾真：形態素クラスタリングによる形態素解析精度の向上,自然言語処理Vol.5,No.2,(1998)

10）白井清昭，徳永健伸，田中穂積：括弧付きコーパスからの日本語確率文脈自由文法の自動抽出,自然言語処理Vol. 4, No. 1,(1997)

11）http://www.iijnet.or.jp/edr/IPA-seika.html

12）Timothy Baldwin, 徳永健伸,田中穂積「動詞多義性解消における語彙交替現象」情報処理学会自然言語処理研究会126-7,(1998)

13）T.Ogino, H.Miyoshi, F.Nishino, M.Kobayashi, "An Experiment on Matching EDR Concept Classification Dictionary with WordNet" IJCAI-97 Workshop on Ontologies and Multilingual NLP, (1997)

14）藤本正樹,桧山努：日本語情報海外発信促進のための言語知識コンテンツ蓄積・流通ソフトの開発,創造的ソフトウェア育成事業及びエレクトロニック・コマース推進事業中間成果発表論文集,IPA,(1997)

15）荻野孝野,小林正博：日本電子化辞書における概念体系,第５回国立国語研究所国際シンポジウム, (1997)

16）大石 亨, 松本 裕治：格パターン分析に基づく動詞の語彙知識獲得, 情処理学会論文誌 第36巻第11号,(1995)

17）Piek Vossen, Wim Peters, Pedro DIez-Orzas,Multilingul design of Euro WordNet, IJCAI-97 Workshop on Ontologies and Multilingual NLP,(1997)