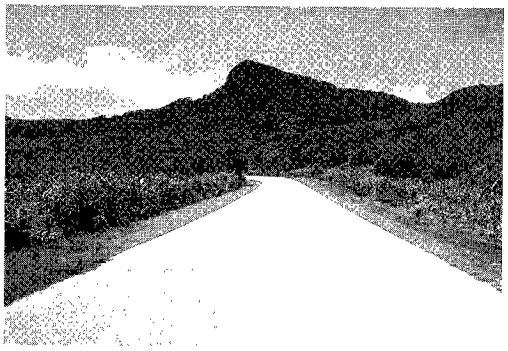
道しるべ： 機械翻訳技術とその適用



大山　芳史
日本電信電話（株）
コミュニケーション科学基礎研究所

インターネットの発展で，世界中の情報が瞬時に取得できるようになり，計算機による自動翻訳のニーズは高まっている．機械（計算機）による，日常使う言葉（自然言語）の扱いが難しい点を，身近な例で説明し，それぞれの課題を克服する技術や歴史について概観する．また機械翻訳に必要な電子化辞書についても紹介する．

■はじめに

雨が一生懸命に降っている

この文は昔作った英日翻訳システムの出力である．英和辞書には“hard”は「一生懸命に，激しく，ひどく」などの訳語が列挙されている．“It's raining hard.”を翻訳すると，場合によりこの訳文が出てくる．「一生懸命に」という表現は，人など主体となるものが主語になるときに使えるがそれ以外では使えない．このことが計算機に分かっていると，「雨が激しく降っている．」と正しく翻訳できる．

（水道から水が出ているのを見て）水，消して

この文は私の娘が幼児のときに話した言葉である．ガスがついているとき「ガス，消して／止めて」は，両方使えるが，「水，消して／止めて」では，「消して」は，使えない．機械翻訳（MT：Machine Translation）でも，語の持つ意味を捉えることと，的確な文章として生成できることが重要なポイントとなる．

科学技術文献の翻訳実験で，“He is a boy.”を翻訳したら，「ヘリウムは少年です」となった，と文献1）で報告されている．1つの単語に1つの訳語で対応できる場合には問題は少ないが，専門用語や固有名詞など単語を追加していくと，he（彼）とHe（ヘリウム），日本語であれば，平野（へいや）と平野（ひらの）のように，複数ある語義や訳語をどのようにして自動的に選択できるかという問題が出てくる．機械翻訳に限らず日常の言葉を扱う自然言語処理では，さまざまなレベルでの多義との勝負になってくる．

以下では，機械翻訳についての歩みを概略し，この分野を研究する際に知っておくと有益な文献，URL，会議などを紹介する．

■機械翻訳の研究の経緯

機械翻訳は，計算機の開発とともに始まり，'50年代には米国でロシア語から英語への簡単な機械翻訳システムの実験がなされた．日本でも九州大学と通産省電気試験所（現在の電総研）で研究が開始された．

'60年代に入り，欧州をはじめ各国で機械翻訳の研究が行われていたが，'66年に米国で，機械翻訳の効率は人手翻訳に比べて大きく劣るため基礎的な研究を推奨するというALPACレポートが出されたのを契機に，米国では言語学の基礎研究を行う方向へ転換し，機械翻訳の研究は冬の時代となった．しかしながら，商用システムとしてのSYSTRAN（当初，露仏翻訳）がECで利用され，一方，カナダでは天気予報を前編集や後編集せずに完全自動で英仏翻訳するシステムTAUM METEOが'76年に実用化された．

日本では，科学技術論文の抄録の翻訳のために科学技術庁のMuプロジェクトが京都大学を中心にして'82年に始まり，構文トランスファー方式による研究が4年間継続された．その後，JICSTの実用化システムへと技術継承された．このプロジェクトを契機に，国内のメーカから日英／英日の機械翻訳システムが研究開発され，商品化に至っている．これらの詳しい経緯は，文献1）～4）に紹介されている．

一方，アジア圏の言語を対象に，ODAの一環として通産省の主導で中間言語方式による多言語機械翻訳システム開発の国際プロジェクト（CICC）が，'87～'94年に実施された☆1.

これら翻訳プログラムの研究と同期して，'86年にEDR電子化辞書プロジェクト☆2が開始され，9年間にわたり開発が行われた．また，'86年より情報処理振興事業協会（IPA）から，「計算機用日本語基本辞書IPAL☆3」が公開され，利用できるに至っている．

NTTでは，新聞記事を対象に，計算機による音声出力システムの研究を'81年に開始した．この過程で構築した43万語の実用的語彙規模の辞書と形態素解析技術をもとに，'85年に日英機械翻訳の研究を開始し，産業経済記事など記述文を対象に意味解析型の日英機械翻訳システムALT-J/E[6] ☆4を実現し，'97年9月にALT-J/Eの意味辞書の一部を「日本語語彙大系（全五巻）[5] ☆5」として岩波書店より出版した．音声を対象とした研究は，ATRが，'86年から国際会議やホテルの予約など場面を限定した音声対話の翻訳の実験[19]に着手し，現在日英独韓中の言語を対象に，リアルタイムで翻訳する実験システムの研究を継続中である☆6．NHKでは'89年から英文のTVニュースや外電ニュースを対象に英日翻訳の試用が行われてきた[7].

■機械翻訳システムの方式

機械翻訳システムには，大きく分けてトランスファー方式，中間言語方式がある（**図-1**）．トランスファー方式は解析，変換，生成のステップを持ち，中間言語方式は，言語に依存しない中間言語への解析とそこから目的の言語への生成の2ステップで翻訳する方式である．中間言語方式では，一般に対象言語数（n）が増えても変換ステップがなく，n種類の解析と生成機能を持てば良いわけであるが，言語が増えるたびに，本来変わらないはずの中間言語仕様が影響を受ける恐れがある．一方トランスファー方式は，言語間の意味体系や構文体系が類似している場合は，解析を比較的浅くして（さぼって）翻訳することも可能となる．もちろんn言語間の翻訳ではn×（n-1）種類の変換ステップが必要となる．これらの方式は，人手によりルールを構築する場合が多く，そういう意味でルールベース翻訳と呼ばれる．

これに対して，既存の用例を用いて翻訳を行う用例ベース翻訳方式がある．これは，原言語と目的言語のペアを蓄積しておき，その中から同一または類似した文章を用いて高品質な翻訳をねらう方式である．用例ベース翻訳では，いかにたくさんの用例対を準備できるかがカギである．日本語と英語の対応付けられた言語データは，欧米圏の言語間に比べて，まだ少ない．

このほか高品質をねらう翻訳としては，文章の可変部分を変数とするテンプレート翻訳も有効で，天気予報，決算速報[24]，特許など特定分野に適用し，実用化されている．この場合，どのようなパターンをテンプレートとするか自動的に定型パターンを抽出する技術も重要で，各種言語に適用できるn-gram統計処理が研究されている．これら各方式を組み合わせたハイブリッド型の翻訳方式は，各方式の長所を引き出すことで，実用性を指向した有力な方式と考えられる．

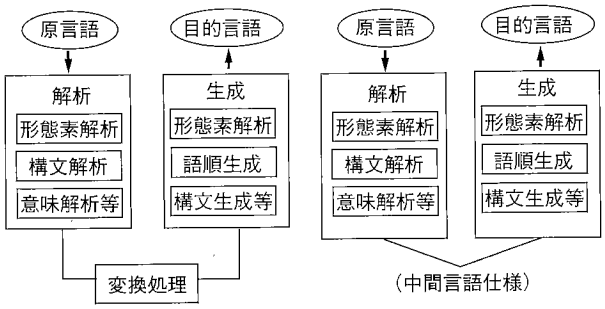


図-1 機械翻訳の方式イメージ

一方，利用される形態でみると，完全自動翻訳および翻訳者の支援としての位置付けがある．翻訳支援の形態としては，改版されるマニュアルの翻訳があり，適宜専門用語を登録したり，既発行済みの対訳データベースを用いることで，翻訳の効率化が図れる．

■機械翻訳を実現する技術

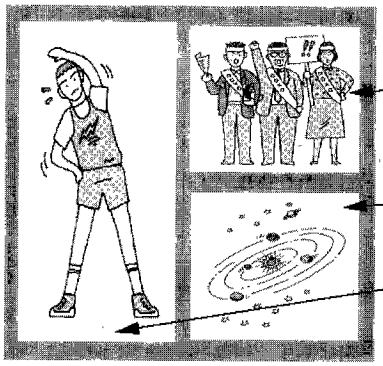
機械翻訳の要素技術としては，形態素解析，構文解析，意味解析，生成，およびこれらを効率よく実現するための支援技術や翻訳文評価技術などがある．

• 形態素解析

形態素解析は，テキストを単語単位に区切り，品詞・活用情報を付与し，場合により，音韻情報（読み等），意味情報，その他語義の識別情報等を付与する処理である．英文では，単語間にスペースがあるため，全体としては処理は簡単である．一方，日本語は，分かち書きする習慣がなく，名詞が連続した複合語を形成しやすいため，その正確な分割処理が必要となる．

日本語の形態素解析は，機械翻訳以外の情報検索やメディア変換でも利用される共通的な技術で，大学など各種研究機関で実用的な語彙規模で実現され公開されている．出力される形態素の単位や品詞体系は，システムによって，必ずしも同一ではない．たとえば，「によって」や「取り立て」を1つの形態素とするものや，「に／よっ／て」や「取り／立て」のように細かく分割するものがある．翻訳する場合，前者のように長めで機能的な分類を用いる

☆1 http://www.cicc.or.jp/homepage/japanese/public/about/act/mt/mt.htm
☆2 http://www.iijnet.or.jp/edr/J_index.html
☆3 http://www.ipa.go.jp/STC/NIHONGO/IPAL/ipal.html
☆4 http://www.kecl.ntt.co.jp/icl/mtg/
☆5 http://www.iwanami.co.jp/hotnews/GoiTaikei
☆6 http://www.itl.atr.co.jp/matrix/



運動する（うんどうする）
（1）16行動　（動作　受身不可）
N1がN2に運動する
N1 campaign for / against N2
[N1（3主体）
N2（3主体 1001抽象物 1236人間活動）]
（2）18物理的移動　（動作　受身不可）
N1が運動する　N1 move
[N1（527天体 706無生物）]
（3）23身体動作　（動作　受身不可）
N1が運動する　N1 exercise
[N1（4人 535動物）]

図-2　「運動する」の語義（日本語語彙大系[5]の場合）

方が，後段の処理が容易になる．

また，OCRや音声等の入力を前提とする場合は，誤りがある前提で処理をする必要があるため，辞書や統計情報により，複数の解を積極的に出力したり，OCRからの複数候補を含む文字列を効率よく解析する機能を備えているのもある[8]．

公開されていてWWWで直接利用可能な形態素解析ツール「JTAG☆8」，プログラムを提供しているもの「茶筌☆9」，「すもも☆10」，「ALTJAWS☆11」，「Breakfast☆12」，両形態のもの「JUMAN☆13」がある．ALTJAWSは研究機関に提供してきたが，日本語語彙大系の意味情報を付与した版の提供を検討中である．このほかにも，英文の品詞を付与するBrill's taggerなど世界中から各種ツールが提供されている．

• 構文解析

構文解析は，文法の規則に基づいて，文章の構造を解析する処理である．日本文は，比較的語順が自由に入れ替わったり，省略ができるため，係り受け解析を用いるのも多い．係り受け解析は，どの語句がどの語句にかかるかを自動的に解析するもので，辞書の単語（語彙）に解析で必要となる情報を収録して解析する方法，係り受け正解データのコーパス等をもとに統計的な手法で解析する方法がある．前者のうち長文に対しても精度が上がるように従属節の分類を詳細に行い解析している手法[9]，後者で，機械学習の手法により決定木（decision tree）を構築しそれにより解析を行う手法[10]がある．英文の構文解析では，語順の制約が記述しやすい句構造規則を用いる場合が多く，この手法をはじめ各種解析のアルゴリズムが，文献2），25）で紹介されている．公開されている代表的なプログラムには，京都大学のKNP，リコーのQJPがある．各種解析プログラムに関して徳島大学工学部知能情報工学科北研究室☆14からのリンクが充実している．

• 意味解析

単語が使われている語義（たとえば「運動する」は，「物体が位置を変える」，「身体を動かす」等の意味がある（図-2）．）を決め，語句間の意味的な関係を解析する技術である．トランスファー方式では，この解析を行う段階で英語側の語彙や構文構造が決定されていく．「天体が運動する→move」「選手が運動する→exercise」などである．また，日本語では頻繁に省略される語句を補完したり，照応（語の指示する対象を認定）したりする文脈処理も，意味を解析して行われる場合が多い．

• 生成

英語の生成では，文章構造（複文か単文か，句にするか文にするか）や，時制（過去，現在など）を決めたり，修飾句の範囲の決定（並列句の部分の束ね方をどうするか），語順の決定，名詞の数（単複）の決定，所有代名詞の生成（妻→my wife），助数詞（枚：2枚の手紙→two letters／a two page letter）の処理，さらに代名詞化をするかどうか，句の省略をするかどうかなどがある．さらに修辞的な考慮や，音声で出力する場合は聞き取りやすい語句を生成したり，速報ニュースのような場合は短い単語へ書き換えるなど，適用先によって必要となる処理もある．

また，これらの処理のステップは，曖昧さを残しながら次のステップへ渡すとか，バックトラックするなど各ステップ間で連携した処理が必要となったり，意味的な知識を辞書にどこまで確定して持たせられるか，前後の文の解析で得た知識をどのように使うかなど，システム化にはさまざまな課題がある．

■辞書・コーパスの開発の動向

電子化辞書など計算機が直接扱える言語知識は，海外国内を含めて最近多く公開されてきている．形態素解析や構文解析に利用できる辞書として，EDR辞書[11]，IPAL辞書[12]がある．EDR辞書は，約40万の概念について約6,000の分類がなされている概念辞書の他，単語辞書，対訳辞書，共起辞書，専門用語辞書，およびEDRコーパス（日本文：22万文，英文：16万文）からなる．IPAL辞書は，基本的な動詞（861語），形容詞（136語），名詞（1,081語）について，形態，意味，統語，慣用表現などに関して詳細に記載されている．

☆8 http://lambda.cipl.cae.ntt.co.jp/jtag/
☆9 http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html
☆10 http://www.brl.ntt.co.jp/sumomo/
☆11 http://www.kecl.ntt.co.jp/icl/mtg/resources/index-j.html
☆12 http://www.fujitsu.co.jp/hypertext/free/breakfast/license.html
☆13 http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html
☆14 http://www-a2k.is.tokushima-u.ac.jp

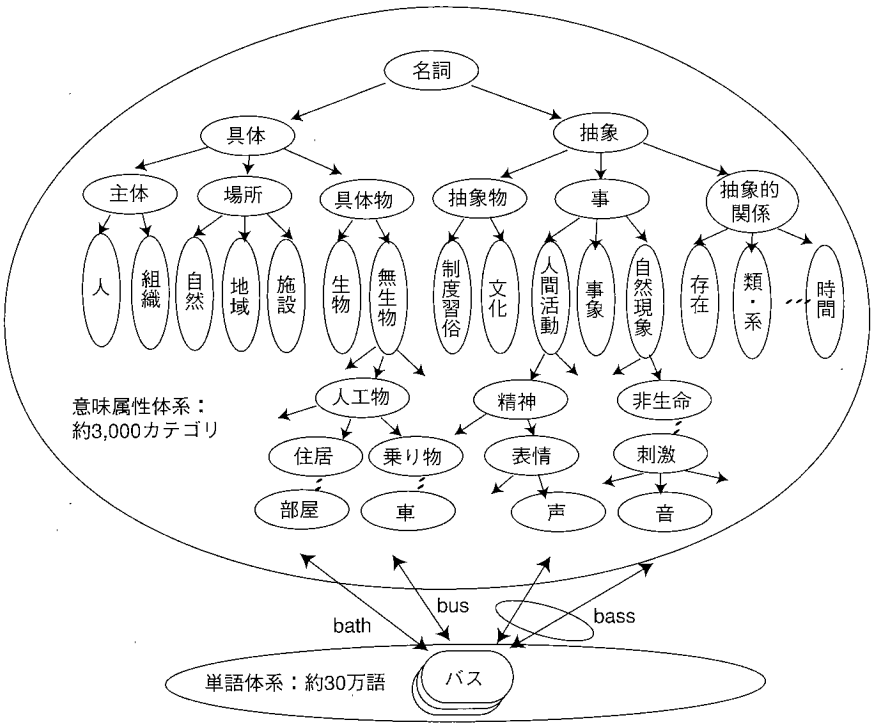


図-3　意味属性体系と単語の関係[5]（バスの例）

また，語彙の体系を記述したものとしては，分類語彙表[13]，角川類語新辞典[14]があり，電子媒体でも提供されている．海外では古く1852年にRogetのシソーラス[15]が出版され，改版されてきている．日本語語彙大系は，約3,000のカテゴリからなる意味体系，カテゴリ情報と品詞情報等を付与した約30万語の単語体系，約6,000語の日本語用言に対して取り得る日本語の構文パターンと英語の構文パターンをペアとして約16,000ペアの構文パターンを収録した構文体系から構成される．先に示した図-2は，日本語語彙大系において，動詞「運動する」に対する3個の構文パターンを定義している例，**図-3**は，「バス」に対する意味体系の対応例である．

米国のプリンストン大学で'85年から開発されたWordNet[16] ☆15は英単語に対して名詞と名詞，動詞と動詞の関係の体系化がなされ，同義・反義等の関係が記述されて，WWWで公開されている．欧州でEuro WordNetとして多言語化がすすめられ，EDR概念体系とWordNetとの照合も試みられている[20]．

また，言語データを集めたコンソーシアムとして，'92年に米国で各組織が保有する言語知識データを共有するためのコンソーシアムLDC☆16が設立され，各種言語データが提供されている．プレーンなテキストデータだけでなく，人手で解析した情報付きの言語データも数多く提供されている．日本でも，日本語の言語データは，新聞社から新聞記事CD-ROMが，提供されている．日本電子工業振興会では，環境白書や経済白書などの日英データを文対応付きで作成しており[17]，今後品詞情報など多様なデータが付与される計画がある．言語データの各種利用可能な情報が，大学等のWWWにまとめられている☆21〜☆23．

■言語データの利用

言語データが電子化され，さまざまなレベルで言語データに正解情報が付与されるようになってくると，自然言語処理プログラムの品質の客観的な評価に利用されたり（たとえば，ルールベースのどのルールが重要かの検証も比較的容易になる），統計や機械学習の手法で，解析プログラムが構築できる[10]．最近では，EDRのコーパスやWordNetを使った研究報告も多く見受けられる．

かけ込み乗車は（危険ですから），おやめください．
For your safety, don't rush your train.

この文章は，JRの新幹線の座席のテーブルに記載されている．対訳で表現されているが，「危険ですから」に該当しそうな部分は英語では逆の意味の“For your safety”となっている．また，“train”は，乗車の「車」になっていて，単語対応のつけにくい対訳例であるが，差異として，このような違いを取り出すことができれば，言語の教育用のツールとしても有用になると思われる．

■翻訳評価技術

翻訳システムの評価には，評価体系が重要となる．NTTの評価試験文は，原言語（日本語）の性質と表現の種類，

☆15 http://www.cogsci.princeton.edu/~wn/
☆16 http://www.ldc.upenn.edu/
☆21 http://www-nagao.kuee.kyoto-u.ac.jp/
☆22 http://tanaka-www.cs.titech.ac.jp/
☆23 http://cactus.aist-nara.ac.jp/lab/resource/resource.html#KYOKI

および原言語と目的言語（英語）の相違点を整理して，日本語側の表現体系をベースに試験項目を体系化し整理し[22]，その一部を公開している（日英3,700文対）☆17．また，JEIDAでは，客観的かつ実用的に評価が行えることをねらいとして，英日（約770文），日英（約400文）の試験文（テストセット）を構築し公開している[21] ☆18．

■機械翻訳の適用例と商品

機械翻訳は実フィールドでは，パソコン通信における翻訳サービスやマニュアルの翻訳者の支援として原稿の下訳作成に使われるケースがある．速報型としては，決算速報の自動英訳サービスへの適用されているALTFLASH[24]がある．これは日本語記事を受信して，人手を介さずに完全自動で速報の見出し文を1秒以内で英訳するもので，'98年の3月から利用されている．市販されている翻訳ソフトは，AAMTジャーナル☆7で日本のメーカを中心に毎年一覧がレポートされている．'98年7月現在，21社で100種類を超える製品が紹介されている．日英・英日が中心であるが，日韓・韓日や英仏独伊西葡→日の製品もある．

■参考書・会議・研究会

教科書としては，機械翻訳を含め自然言語処理に関して文献2)，解析技術を中心として文献25)，電子辞書については文献18）が参考になる．機械翻訳の解説には文献23)，機械翻訳の歴史からシステム紹介までは文献1）～4）がある．また，年4回発行のAAMTジャーナルでは，毎号「技術早分かり」のコーナで機械翻訳関連の技術（'99年1月号では，省略要素補完技術）が紹介されている．

国際会議は，COLING，ACLの他，主に機械翻訳に関する国際会議としてTMI，MT-SUMMITがある．TMIは機械翻訳に関するアプローチや技術について議論がなされている．次回TMI-99は，8月にチェスター（英国）☆19で開催される．MT-SUMMITは，アジア，米国，欧州と持ち回りで隔年で開催されている．機械翻訳の研究者だけでなく，MTシステム販売で商売をしている人や，MTを使って翻訳に利用している人も集まる機械翻訳ビジネスとしての情報の交換の場でもある．次回MT-SUMMIT99☆20がシンガポールで開催される．また，環太平洋の国々を中心として，自然言語処理に関する国際会議PACLINGやNLPRSも開催されている．また，ACLがスポンサのWVLC（Work Shop on Very Large Corpora）は，'97年で5回目を迎え，各種コーパスを使った言語処理に関して発表されている．

関連する国内の研究会を以下に示す．言語処理学会NLP（http://www.crl.go.jp/pub/nlp/）は，3月に年次大会が開催され，今年は約140件の発表が予定されている．

【情報処理学会】
- 知能と複雑系研究会（ICS）

http://www.ymd.dis.titech.ac.jp/sig-ics/
- 自然言語処理研究会（NL）

http://cactus.aist-nara.ac.jp/staff/utsuro/SIGNL
- 音声言語情報処理研究会（SLP）

http://tk01.tk.elec.waseda.ac.jp/~koba/SLP/

【電子情報通信学会】
- 思考と言語研究会（TL）

http://www.pluto.ai.kyutech.ac.jp/TL/
- 言語理解とコミュニケーション研究会（NLC）

http://www.ieice.or.jp/iss/jpn/nlc/nlc-index-j.html

【人工知能学会】
- 言語・音声理解と対話処理研究会（SIG-SLUD）

http://winnie.kuis.kyoto-u.ac.jp/sig-slud/
- ことば工学研究会（SIG-LSE）

http://www.kecl.ntt.co.jp/banana/Workshop/

参考文献
1）長尾　真: 機械翻訳はどこまで可能か, 岩波書店 (1986).
2）長尾　真（編）: 自然言語処理, 岩波書店 (1996).
3）Whitelock, P. and Kilby, K.: Linguistic and Computational Techniques in Machine Translation System Design 2nd Edition, UCL Press, London (1995).
4）成田　一: パソコン翻訳の世界, 講談社現代新書 (1997).
5）池原　悟, 宮崎正弘, 白井　諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林　良彦（編）; 日本語語彙大系, 岩波書店 (1997).
6）八巻俊文, 大山芳史, 白井　諭, 横尾昭男: 機械翻訳特集, 日英機械翻訳システムALT-J/Eの研究開発, NTT R&D, Vol.46, No.12, pp.1391-1398 (1997).
7）相沢輝昭, 加藤直人, 鎌田雅子: 外電経済ニュースの英日機械翻訳, 情報処理学会論文誌, Vol.37, No.6, pp.1041-1048 (June 1996).
8）Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, Proc. 15th COLING, pp.201-207 (1994).
9）白井　諭, 池原　悟, 横尾昭男, 木村淳子: 階層的認識構造に着目した日本語係り受け解析の方法とその精度, 情報処理学会論文誌, Vol.36, No.10, pp.2353-2361 (Oct. 1995).
10）春野雅彦, 白井　諭, 大山芳史: 決定木を用いた日本語係受け解析, 情報処理学会論文誌, Vol.39, No.12, pp.3177-3186 (Dec. 1998).
11）荻野孝野: EDR電子化辞書について, 情報処理学会情報メディア研究会, Vol.34, No.7, pp.31-38 (1998).
12）情報処理振興事業協会技術センター: 計算機用日本語基本動詞辞書 IPAL, 解説編＆辞書編 (1987).
13）国立国語研究所: 分類語彙表, 秀英出版 (1964).
14）大野　晋, 浜西正人: 角川類語新辞典, 角川書店 (1981).
15）Lloyd, S. M.: Roget's Thesaurus, Longman Group Ltd. (1982).
16）Fellbaum, C.: WordNet an Electronic Lexical Database, MIT Press (1998).
17）（社）日本電子工業振興協会: 自然言語処理システムの動向に関する調査報告書 (1997).
18）Wilks, Y. A., Slatir, B. M. and Guthrie, L. M.: ELECTRIC WORDS, MIT Press (1996).
19）ATR国際電気通信基礎技術研究所（編）: 自動翻訳電話, オーム社 (1994).
20）（財）日本情報処理開発協会: 大規模知識ベースに関する調査研究－オントロジー工学に関する調査研究－報告書 (Mar. 1998).
21）（社）日本電子工業振興協会: 機械翻訳システム評価基準 (1995).
22）池原　悟, 白井　諭, 小倉健太郎: 言語表現体系の違いに着目した日英機械翻訳機能試験文項目の構成, 人工知能学会誌, Vol.9, No.4, pp.569-579 (1994).
23）田中穂積: 機械翻訳の過去・現在そして未来, 電子情報通信学会誌, Vol.78, No.11, pp.1171-1176 (1995).
24）内野　一, 大山芳史: 速報型日英翻訳システムALTFLASH, NTT技術ジャーナル, Vol.11, No.3, pp.81-83 (1999).
25）野村浩郷: 自然言語処理の基礎技術, 電子情報通信学会編 (1988).

（平成11年3月11日受付）

☆7 http://www.jeida.or.jp/aamt/
☆17 http://www.kecl.ntt.co.jp/icl/mtg/resources/
☆18 http://www-karc.crl.go.jp/ips/jeida/document.html
☆19 http://www.ccl.umist.ac.uk/events/tmi99/
☆20 http://www.crl.go.jp/pub/nlp/CFP/MT-Summit99