自 然 言 語 処 理 101－3
（1994．5．27）

相関表現を含む英語長文の文構造解析

田添 丈博* 相馬 陽一** 河合 敦夫** 椎野 努** 桝井 文人*** 杉尾 俊之***

*鈴鹿工業高等専門学校 電子情報工学科
**三重大学 工学部
***沖電気工業 (株) 関西総合研究所

要旨

一般に、英語長文には、省略現象・並列現象・慣用的表現など、人間には理解できても、既存の辞書や文法ルールでは処理が困難な言語現象が多数存在する。これらの言語現象は機械翻訳において大きな問題点となっている。本稿では、従来手法では処理困難な言語現象として相関表現を対象にし、文構造決定のための新しい解析手法を考案し、合わせて省略現象・並列現象の解析へのアプローチを試みた。解析アルゴリズムは、相関表現において連結される2つの要素に対して、キーワード推定・スコープ推定・バックトラックの3つのプロセスから構成されている。また、英字新聞からの130文を適用対象とし、本解析アルゴリズムの有効性を検証した。

A Method of Parsing Correlative Structure in Long English Sentences

Takehiro Tazoe* Youichi Souma** Atsuo Kawai** Tsutomu Shiino**
Fumito Masui*** Toshiyuki Sugio***

* Electronics and Information Engineering, Suzuka College of Technology
** Faculty of Engineering, Mie University
*** Kansai Laboratory, Oki Electric Industry Co.,Ltd.

Abstract

Generally, there are many phenomena of language (ellipsis, paratqaxis, ideomatic expression and others), which are easy to understand for men but are difficult for existing language processing systems. Parsing these phenomena is one of the important problems in machine transration. We analyzed correlative structure, which has ellipsis and paratqaxis at the same time, and designed a new method of parsing it. This parsing algorithm consists of three processes; assumption of the key word, assumption of the scope and backtracking. We evaluated effectiveness of this parsing algorithm with 130 sentences in English newspaper.

1 はじめに

これまでの自然言語処理の研究は、機械翻訳システムを始め、個々のシステムの持つ辞書や文法ルールによって明確に定義される文を対象に行なわれ、その技術が蓄積されてきた。しかし、実際の日常文には省略現象、並列現象、慣用的表現など、人間には理解できても、既存の辞書や文法ルールでは処理が困難な言語現象が多数存在する。今後、実用的な自然言語処理を目指すには、これら日常使用される従来手法では処理困難な言語現象にも柔軟に対処できる技術が必要となってくる。一般に英文においては、これらの問題となる言語現象は長文中に顕著に見られ、長文解析の精度向上のための重要な問題点の1つとなっている。

そこで、前述のような問題となる言語現象を比較的多く含む表現として、相関表現を取り上げることにする。相関表現とは、「連結詞が分離し、前後相応じて1個の連結詞としての役目を果たす」統語構造であり、連結される2つの要素には並列現象、省略現象が見られる。我々は、統語的、統計的側面からのアプローチによって相関表現を含む長文の文構造を決定する手法を考案した。

具体的には、連結される2つの要素の単語列の情報を用いて各々の要素の中心となるキーワードを推定し、そのキーワードに基づいて各々の要素の範囲(スコープ)を推定する手法である。また、スコープ推定に失敗した場合は、バックトラックがかかりキーワードを推定しなおすようになっている。この手法によって90%以上の割合で相関表現を含む長文の正しい文構造の決定に成功した。このことによって、長文に対してより深い意味の理解が可能となり、英日機械翻訳においては、より的確な日本語訳出が可能となると考えられる。

本稿では相関表現の特徴をとらえ、それをもとに解析する手法と、実用例文に対するルール適用結果について述べる。また、この実験結果をもとに解析アルゴリズムを検証する。

2 相関表現の特徴

2.1 統語構造

相関表現とは、「連結詞が分離し、前後相応じて1個の連結詞としての役目を果たす」統語構造である。具体的には、as...as, more...than, so...as, rather...than, not only...but, less...than, either...orなどが挙げられる。

相関表現には特定のマーカとなるべき単語(as...asなど)が必ず存在するため、相関表現であるかどうかの識別は容易である。ここで、連結される2つの要素とマーカとの位置関係から、相関表現を図1のように2つのタイプに分類することができる。一般的に、タイプ1は比較構文となり、タイプ2は等位並列となる。

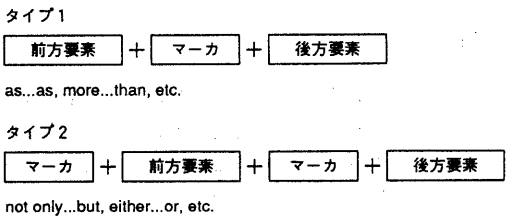

図 1: 相関表現の分類

本稿では比較構文(タイプ1)を中心に論じる。なお、比較要素の中心となる単語をキーワードと呼ぶことにする。

比較構文には、比較特性および2つの比較要素が必ず存在する。比較特性には、形容詞句、副詞句、形容詞を伴う名詞句などが相当し、2つの比較要素は構文上・意味上同質のものが相当するという特徴をもつ。

この比較要素(キーワードおよびそのスコープ)が推定できれば、図2のような変形操作も可能となり、比較構文を簡単な2文に分割することによって解析が容易になると考えられる。

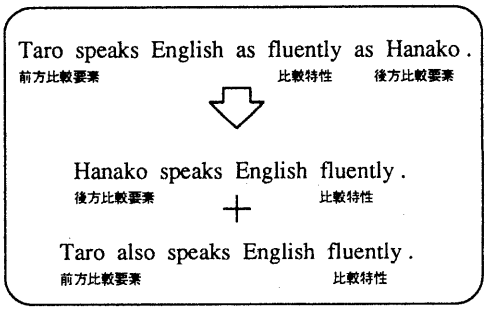

図 2: 比較構文の変形操作

具体的には、2つ目のasで2文に分割し、TaroとHanakoが比較要素であることが認識できれば、後文Hanakoの後ろに前文からTaroを除いたspeaks Englishおよび比較特性であるfluentlyを補完し、1つ目のasを副詞alsoに置換して適当な位置に移動させればよい。

また、この変形操作から、比較構文において、2つの比較要素が並列現象のような振舞いをし、省略現象を含んでいるということがわかる。

この、比較構文における後方の文の省略は大別して、

1. 主語以外を省略
2. 主語＋(助)動詞以降を省略
3. 主語＋(助)動詞を省略
4. 比較特性のみを省略

に分類することができる。いずれも、比較特性をはさみ前後の共通部分の反復を避けるための規則である。

このように、2つの比較要素の認識といずれの省略規則が使われているかという認識ができれば、図2のような変形操作が可能になる。一方、比較要素はこの例のように単語だけという場合は少なく、冠詞や形容詞による前置修飾や、前置詞句や関係代名詞節による後置修飾を伴うのが一般的である。よって、比較要素を決定するためには、キーワードの推定と同時にこれら修飾語句も含めたスコープを推定する必要がある。

2.2 出現頻度

実用例文として「Japan Times」の6カ月分(46489文)を用いて、比較構文の頻度調査を行なったデータを表1に示す。ここで、長文は30語以上の単語から構成される文と定義している。

表1: 比較構文の出現頻度

	対全文[%]	対長文[%]
同等の比較	2.9	6.7
差異の比較	3.2	5.7
合計	6.1	12.4

この表から、比較構文が長文となりやすいことがわかる。

同等の比較は、as...as構文、so...as構文、such...as構文などよりなるが、全体の90%以上はas...as構文であり、またas...as構文のうち70%程度はas well asなどの慣用表現であった。しかし、一部の慣用表現(as soon asなど)を除き、2.1節で述べた統語上の特徴を持っているため、一般表現と区別する必要はない。

一方、差異の比較は、比較級＋than構文、more ...than構文、rather...than構文、less...than構文など多種に渡るが、thanをマーカとして容易に抽出できる。また、統語上の特徴も同等の比較とほぼ同じであるので、解析には共通の考え方を採用することができる。

3 従来システムの問題点

図2のような英文を従来型の機械翻訳システムに入力すると、「Taroは、Hanakoと同じくらい流暢に英語を話す。」と訳出する。また、HanakoをFrenchに置き換えると、「Taroは、フランス語と同じくらい流暢に英語を話す。」と訳出する。実際には、Taro-Hanako, English-Frenchがそれぞれ意味的に対応するが、従来システムでは何と何が比較されているのかを認識しないため、ともに同じ解析がなされ、同じ形の訳文が出力される。このように、出力文を曖昧にし、読み手に対応関係を決定させるのは機械翻訳においては有効な手法である。しかし、談話理解システムなどの文の内容を理解する必要のある分野への応用性がない。

また、比較要素を決定せず、部分文法と頻度情報の組合せで解析を行なうなどの理由によって、長文においては後文の範囲の推定に失敗する場合が非常に多い。

そこで、第2章で述べた特徴をもとに、前後キーワードを推定し、2つの比較要素のスコープを推定すればよいと考えられる。

4 解析アルゴリズム

4.1 解析処理の概要

処理概要を図3に示す。構文解析部において、句のまとめ上げ処理が終了した後にルールが適用さ

れ、その結果は節のまとめ上げ処理へと渡される。

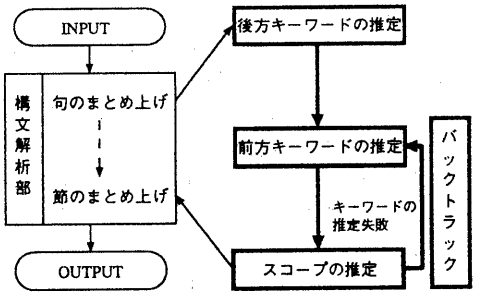

図 3: 解析処理の概要

まず初めに、後方キーワードを推定し、次にこの推定された後方キーワードをもとに前方キーワードを推定する。前後キーワードを推定できたら、推定されたそれぞれのキーワードを中心にスコープを推定する。スコープ推定時に、キーワードの推定に失敗していることを認識したらキーワード推定部にバックトラックがかかる。

句のまとめ上げ後にルールを適用すると、冠詞や形容詞などの簡単な前置修飾部が被修飾単語の下位構造となり、キーワード推定時に余計な単語をキーワード候補とすることが少なくなる。また、文の全体的な構造も把握しやすくなる。

以下では、図 3における各処理部 (太線枠の処理部) について詳しく解説する。

4.2 キーワードの推定

まず初めに、後方キーワードを推定する。本稿においては句のまとめ上げ後にルールを適用しているため、前述のように前置修飾部を意識しなくてもよい。そこで、後方キーワードは、マーカから文末方向に探索し、最初に出現した単語である、と一意に決定することができる。

それに対し、前方キーワードとマーカとの間には後置修飾部がはさまる場合が多く、その後置修飾部にも前方キーワードの候補となる単語が存在することが多い。そのため、後方キーワード推定ルールのように文頭方向に探索を行ない、最初に出現した単語が前方キーワードであると簡単に決定することはできない。

そこで、2 つの比較要素は構文上・意味上同質のものであるという特徴に着目し、一意に決定された後方キーワードと各前方キーワード候補との相違度を計算し、相対的に最も相違度値の低い単語を前方キーワードと推定するという手法をとる。概念的には図 4のようになる。

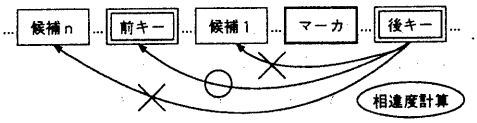

図 4: 前方キーワードの推定

単語間の類似性の尺度となる相違度 Dは以下の式のように、部分相違度 P_iに適当な重み W_iを掛け合わせ、その総和を求めることによって算出する。

$$D = \sum_{i=1}^{n} P_i \cdot W_i$$

部分相違度情報には表 2のような 6 種類の情報を用いる。

表 2: 部分相違度情報

P_n	部分相違度情報	概念レベル		
P_1	前置修飾情報	表層	構文	
P_2	後置修飾情報			
P_3	前置詞句情報			
P_4	距離情報			
P_5	単語構文情報			深層
P_6	単語意味情報			

概念レベルという観点から見れば、上の方に書かれた情報ほど表層的な情報であり、下の方にいくほど深層的な情報である。

以下では、各部分相違度情報および部分相違度の与え方について詳細に解説する。部分相違度を与える範囲は文頭もしくは接続詞のような統語上の切れ目に相当する単語までとした。

(1) 前置修飾情報 : P_1

前置修飾部に同じ単語 (ただし冠詞や限定詞は除く) を持つ単語同士は相違度が低いと考えられるため与える情報。同じ前置修飾部を持つ前方キーワード候補には 1、それ以外には 2 を与える。

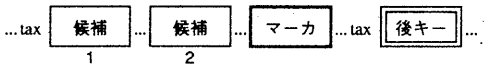

図 5: 部分相違度 P_1の与え方

(2) 後置修飾情報 : P_2

同じ前置詞によって後置修飾されている単語同士は相違度が低いと考えられるため与える情報。同じ前置詞を持つ前方キーワード候補には 1、それ以外には 2 を与える。

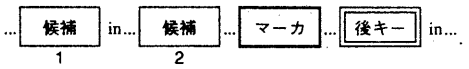

図 6: 部分相違度 P_2の与え方

(3) 前置詞句情報 : P_3

後方キーワードが前置詞句の一部である場合を除き、前方キーワードが前置詞句に含まれることは少ないと考えられるため与える情報。前方キーワード候補が前置詞句の一部であったら 2、それ以外には 1 を与える。

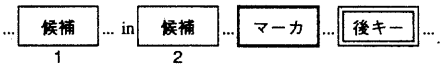

図 7: 部分相違度 P_3の与え方

(4) 距離情報 : P_4

相対的に近くに存在する単語ほど結び付きが強いと考えられるため与える情報。後方キーワードから相対的に距離が近いものから順に 1, 2, …, n を与える。

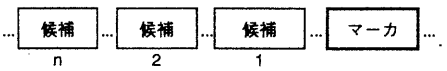

図 8: 部分相違度 P_4の与え方

(5) 単語構文情報 : P_5

単語の持つ構文情報 (格情報など) が似ている単語ほど、単語同士の相違度が低いと考えられるため与える情報。構文情報の一致する割合が高いものから順に、1, 2, …, n を与える。

(6) 単語意味情報 : P_6

単語の持つ意味情報 (階層的分類) が似ている単語ほど、単語同士の相違度が低いと考えられるため与える情報。意味情報の一致する割合が高いものから順に、1, 2, …, n を与える。

4.3 スコープの推定

前後方キーワードを推定できたら、次に、キーワードを中心にそれぞれの比較要素のスコープを推定する。統語的なスコープの切れ目情報、および実用例文を調査することによって得られた切れ目情報を体系化し、この局所的な情報を用いることによってスコープを推定する。これによって、解析部が簡単になり、修飾部に解析困難な言語現象が存在してもあまり意識せず、スコープが決定できるようになる。

具体的には図 9の矢印方向に番号順にスコープを推定していく。

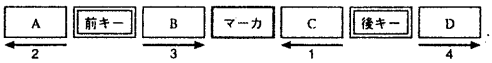

図 9: スコープ推定手順

また、統語上、比較構文のスコープのパターンは図 10の 4 パターンのいずれかに分類することができる。

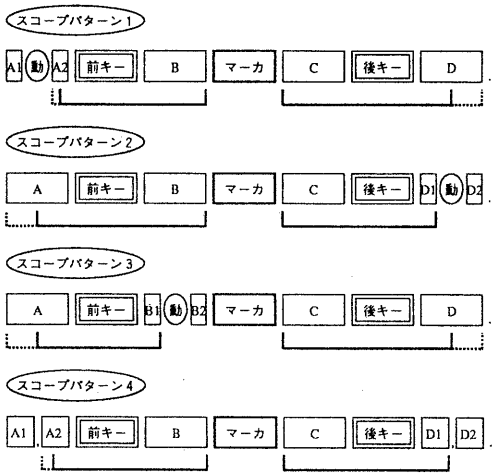

図 10: スコープパターンの分類

このスコープパターンの分類ができれば、スコープの推定も容易になる。しかし、一般に入力文がどのスコープパターンに分類されるかを決定するためには、文の全体的な構造を把握する必要があると考えられる。しかし、本稿では長文を対象にしているため、文の全体的な構造を把握するのは困難である。

そこで、図9に示したような手順で、1つずつ局所的にスコープを推定していき、同時にスコープの切れ目となった情報を保持することによって、図10のいずれのパターンに入力文が相当するのかを限定していき、最終的に全てのスコープを推定する。

もう少しわかりやすく、実際の例文を用いて説明する。例えば、以下のような例文を考える。

(実用例文)

They are also studying the possibility of car production in Europe, so that they can *boost* sales there *beyond* Japan's voluntary [quota] on vehicle exports to the EC [as well as] import [restrictions] set by certain individual European countries.

キーワード推定部によって 前方キーワードがquota、後方キーワードがrestrictionsと推定されている。

まず最初に、後方キーワードの前方スコープを推定する。このスコープは図10からも明らかなように、マーカまでと一意に決定することができる。

次に、前方キーワードの前方スコープを推定する。このスコープは、前置詞 beyond までと推定される。前方キーワードの前方スコープが前置詞で切れる場合はさらに前方の単語を探索する。そして、この例文の場合、スコープ推定と同時に、主動詞 boost という情報を保持する。

この時点で、図10から、例文はスコープパターン1に分類される文であることが判明する。そして、自動的に前方キーワードの後方スコープはマーカまで、後方キーワードの後方スコープは文末もしくは文末を表すような情報までと推定することができる。

よって、先の例文においてはアンダーラインで示すようなスコープを推定することができる。

4.4 バックトラック

前述のような方法で前後方キーワードおよびスコープを推定するが、スコープ推定時に図10のいずれのスコープパターンにも分類できないようなキーワードをキーワード推定部が選出する可能性も考えられる。

そこで、もしそのような単語がキーワードとして選出された場合、前方キーワード推定部にバックトラックし、前方キーワードを次候補に入れ換える。バックトラックは、図10のスコープパターンのいずれかに分類されるか、前方キーワード候補がなくなるまで繰り返される。

5 適用実験結果と検証

2.2節の調査統計で用いた実用例文から任意に抜き出した130文に対して以上のようなアルゴリズムの適用実験を行ない、その有効性を検証した。

5.1 適用実験結果

総合結果は、表3のようになった。全体では、130文中119文の解析に成功し、正解率は91.5%を得た。従来システムには、このような比較対象を決定する処理はないため解析結果を比較することは難しいが、この実験に使用した実用例文を従来型機械翻訳システムに入力してみると、比較の部分での構文的な誤解析が約30%程度、その他の誤解析も合わせると50%を越えたことから、本実験の成果を利用すれば解析精度向上が見込める。

表3: アルゴリズム適用結果

	対象	成功	率[%]
キーワード推定	130	120	92.3
スコープ推定	130	129	99.2
バックトラック	10	0	0.0
総合	130	119	91.5

以下では、もう少し詳しくそれぞれの解析アルゴリズムとその適用実験結果について検証する。

5.2 キーワードの推定部の検証

まず、キーワードの推定に使用した部分相違度情報(表2参照)について検証する。それぞれの部分相違度情報を単独で使用した場合の結果と、重みの概念を導入した場合の結果を表4に示す。

表 4: 部分相違度情報と成功率

部分相違度情報	成功率
P_1	70.8
P_2	75.8
P_3	66.1
P_4	66.1
P_5	69.2
P_6	73.1
$\sum P_n W_n$	92.3

この表から、部分相違度情報を単独で使用した場合の成功率が70%前後であるのに対し、重みの概念を導入すると約90%まで成功率が上がっていることから、重みの有効性が明らかである。

4.2節でも述べたように、相違度計算には適当な重みが必要である。そこで、適用実験から求めた重みの組合せについて検証する。

重みは0〜10の範囲で全ての組合せについて相違度計算を行ない、成功率が最も高くなるような組合せを抽出した。組合せは11^6通り存在するが、高い成功率を得た組合せは多少のばらつきはあるものの、ほぼ一意に決定することができた。その大小関係は、以下のようになった。

$$W_1 \geq W_2 > W_3 > W_4 > W_5 \geq W_6$$

この結果から、表層的な情報ほど信頼度が高く、深層的な情報ほど信頼度が低いことがわかる。これは、視覚的に比較対象を推定し、さらに単語の意味など深層的な情報をもとに比較対象を1対に断定する人間の頭の中での働きと似ている。また、深層情報は与え方が難しい、利用形態が難しいなどの要因も考えられる。

表3の結果は、以下のような重みの組合せを使用した結果である。

$$\{W_1, W_2, W_3, W_4, W_5, W_6\} = \{10, 9, 3, 2, 1, 1\}$$

推定失敗の原因は、距離情報にあると考えられる。距離的に非常に遠い単語が比較対象である場合、距離以外の部分相違度値がよほど小さくないと正しく認識できない。しかし、重みの組合わせにはバランスが必要であるため、単純に距離情報の重みを小さくするだけでは、成功率は上がらない。

また、文脈的な知識や世界知識を利用しないと比較対象を決定することができない以下のような文は、必ずしも正しく解析されるとは言えない。

German speaks English as fluently as French.

この文は、Frenchの比較対象がGermanであるのかEnglishであるのかは、文脈情報がなければ人間にも判断できない。本稿における解析アルゴリズムでは、距離情報(P_4)の違いによって、必ず距離の近い方の単語を推定する。

このように、評価者にも比較対象を決定できないような文は今回の適用実験の対象から外した。

5.3 スコープの推定部の検証

スコープ推定部の推定失敗要因にも文脈情報、世界知識が影響している。

表3よりスコープ推定失敗は130文中1文だけである。まず、実際に失敗した実用例文を挙げる。

(実用例文)

…, drawing voices for and against as well as letters threating to *steal* the nugget *if* it was purchased.

この例文はスコープパターン1であると判定され、アンダーラインで示すようなスコープを推定する。

実際には接続詞if以降の節は動詞stealに係る副詞節であるので、ifの手前でスコープを切るのは適当でなく、文末までスコープに含めるのが正しい。しかし、この判断にはifの直後のitの同定が必要であり、同定するためには文脈情報、世界知識が必要である。このように、統語的に判断できない場合のスコープの正しい認識は難しい。

5.4 バックトラック部の検証

表3にあるようにバックトラッキングによってキーワードの推定失敗を補正することはできなかった。その理由を解明するために、キーワード推定に失敗した10文を検証してみたところ、バックトラックの条件が緩いため、キーワードの推定に失敗しても、それなりにスコープを推定してしまうところにあることがわかった。しかし、現在のルールでは、正しいキーワードを推定したが間違っていると判断し、代りに誤ったキーワードを再度推定するということはない。今後、この特徴を失わない範囲で適用条件を厳しくしていく必要がある。

6 まとめ

本稿で述べた相関表現解析アルゴリズムは、統語的な特徴をもとに、実用例文を調査することによって得られた情報を考慮することによって構成されている。従って、今後、調査対象、適用対象を増やした場合、さらに解析の信頼度を上げる新たな情報が加わる可能性はある。また、本解析アルゴリズムとともに文献[7]のような補完技術を併用すれば、高い精度で省略補完が実現でき、翻訳文の訳出精度の向上につながると考えられる。

並列現象への拡張を考えてみると、1対1対応の並列現象にはほぼそのまま本解析アルゴリズムが利用できる。また、複数のものがカンマで並べられている場合、どこまでが並列であるのかといった範囲の推定に関しては、相違度情報や相違度計算式を改善し、値がいくつ以下ならその単語同士は似ているといったしきい値を設定することによって活用が可能であり、広く長文解析に利用することができると考えている。

参考文献

[1] 「英文法総覧」. 安井稔. 開拓社, 1984.

[2] 「実例英文法」. A.J. トムソン, A.V. マーティネット著／江川泰一郎訳注. オックスフォード大学出版局, 1975.

[3] 「分類語彙表」. 国立国語研究所. 秀英出版, 1990.

[4] 「LONGMAN LEXICON OF CONTEMPORARY ENGLISH」. Tom McArthur. Longman Dictionaries Limited, 1992.

[5] 「長い日本文における並列構造の推定」. 黒橋禎夫, 長尾眞. 情処研報 Vol.91, NO96, 1991.

[6] 「英日機械翻訳における名詞句並列を含む長文解析について」. 田添丈博, 相馬陽一, 河合敦夫, 椎野努, 桝井文人, 杉尾俊之. 第7回人工知能学会全国大会論文集, 1993.

[7] 「"as...as" の特性を利用した英日機械翻訳処理の検討」. 桝井文人, 綱島督之, 杉尾俊之, 椎野努, 相馬陽一, 田添丈博. 電子情報通信学会技術研究報告, 1993.

[8] 「比較構文の特性を利用した英日機械翻訳の検討–実用文コーパスにおける as...as 構文のモデル化–」桝井文人, 綱島督之, 杉尾俊之, 椎野努, 相馬陽一, 田添丈博. 電子情報通信学会技術研究報告, 1994.

[9] 「実例に基づく翻訳における複数翻訳例の組合せ利用」. 佐藤理史. 人工知能学会誌, 1991.