情報学基礎 64－8
自然言語処理 145－8
（2001．9．10）

帰納的学習を用いた構文解析手法の旅行用英会話文における性能評価

正富 欣之　荒木 健治　栃内 香次

北海道大学大学院工学研究科
〒060-8628 札幌市北区北13条西8丁目
TEL: 011-706-7389　FAX: 011-709-6277
E-mail: {tome,araki,tochinai}@media.eng.hokudai.ac.jp

あらまし　自然言語処理における構文解析手法に，固定された文法規則を用いるものと用例から得られたデータを用いるものがある．これらの手法における問題点として，前者は対象に依存した言語現象への対処が困難である，後者はデータ量が膨大になる，ということが挙げられる．そこで，実例から帰納的学習を用いて構文解析ルールを獲得し，得られた解析ルールを適用した構文解析を行うことにより，上記の問題を解消することを目指した手法を提案した．本手法の利点としては，会話文や対話文などの非文法的で省略の多い文にも対処でき、学習アルゴリズムは言語に依存していないので，学習データさえあれば同一のアルゴリズム（帰納的学習）でどのような言語にも適用できるということがある．本稿では，上述の手法に基づいたシステムについての概要を述べ，旅行用英会話文の英文と日本文を用いた実験を行い，本手法の言語非依存性について検討する．

キーワード　構文解析，係り受け，帰納的学習，言語非依存性

Evaluation of a Syntactic Analysis Method Using Inductive Learning for the English Travel Conversations

Yoshiyuki Masatomi, Kenji Araki, Koji Tochinai

Graduate School of Engineering, Hokkaido University
N13 W8, Kita-ku, Sapporo, 060-8628 Japan
TEL: +81-11-706-7389　FAX: +81-11-709-6277
E-mail: {tome,araki,tochinai}@media.eng.hokudai.ac.jp

Abstract　In this paper, we propose a syntactic analysis method using inductive learning from examples. In our proposed method, the system acquires the parsing rules using the examples of parsing result. And the system parses Japanese sentences using the acquired parsing rules. We consider that our proposed method can resolve problems of the Rule-based approach and the Example-based approach. Moreover, we consider that this method can be applied to the other languages. We performed the experiment using the English travel conversations. We consider the result of experiment about the ability of our proposed method.

Keywords　syntactic analysis, dependency, inductive learning, independent on language

1 はじめに

近年，インターネットの普及により母国語とは異なる言語に触れる機会が増大している．これにより，多種多様な言語間での人々の交流も増加した．したがって，種々の言語を母国語に変換する機械翻訳や，母国語以外の言語における情報検索などの技術が必要とされている．つまり，ある特定の言語だけに対応するのではなく，多言語に対応できる言語非依存性を備えた自然言語処理手法が非常に重要となっている．

自然言語処理において，構文解析は文を解析する上で非常に重要である．それゆえ，これまで構文解析に関する多くの研究がなされてきた．それらの研究において，主となる構文解析手法は文脈自由文法などに基づく文法規則を用いたものである．文法規則に基づく手法では，予め与えられた規則が固定したものであるため，想定された言語現象に対しては正しく解析される．しかしながら，実際の自然言語処理を行う上では，予め想定し得ないさまざまな言語現象が存在する．したがって，規則を与える手法では，さまざまな言語現象に対処するのが困難である．また，それらの言語現象に対処できるように人手で新たな規則を追加しようとすれば，既存の他の規則との整合性をとるために非常に労力のかかる作業が必要となる．

このようなさまざまな言語現象に対処するために，機械翻訳の研究では用例に基づく手法が試みられている[1, 2, 3]．同様に，構文解析の研究でも用例に基づく手法を用いた構文解析システムが提案されている[4]．一般に，用例に基づく手法は，人間が例を用いて問題を解決する過程を模倣し，大量の例を用い，与えられた入力文と類似度の高い用例の一部を組み合わせることにより解析する手法である．これらの手法では，新たに用例を追加することにより精度が向上し，システムの保守および更新が容易である．しかしながら，このような手法に基づくシステムの性能，つまり構文解析の質を向上させるためには膨大な構文解析例を必要とする．それゆえ，大規模な構文タグ付コーパスが必要となるが，作成には多大な労力がかかる．また，構文解析例を作成するために構文解析ツールを使用する場合には，そのツールの解析精度が問題となる．

また，決定木を用いた構文解析手法も提案されている[5]．この手法では統計情報を用いて構文解析を行う．このように統計情報を用いて規則に重み付けを行う手法では，統計情報を得るために用いるコーパスをどのような基準で選択し，どれだけの量を使用するかという問題が残る．

これらの問題を解決するために，比較的少量の構文解析例を用いて精度の高い構文解析を行える手法として，構文解析例（品詞付単語列と正しい構文解析結果の組）からの帰納的学習を用いた構文解析手法を提案した[6]．構文解析例からより多くの解析ルールを自動的に抽出することができれば，少量の構文解析例からでもさまざまな言語現象に適応できる精度の高い構文解析が可能となる．ここで獲得される解析ルールは言語毎に存在する，例えば日本語文法，英文法などのような文法規則を表すものではなく，文字を記号列と捉えた変換規則ということができる．このように記号列の変換規則を用いることにより言語に依存しない解析ができると考える．本手法で用いる帰納的学習では，用例に基づく手法のように用例の一部分を組み合わせたり，統計的な手法のように確率を利用したりするのではなく，用例中に内在するルールを獲得し，それを用いて解析を行う．本手法に基づくシステムは1文の入力毎に順次学習を繰り返し，入力に対して動的に適応していくので、コーパスの種類についてはどのようなコーパスを選んでも良く、そのコーパスの量は学習が飽和したと考えられる量を入力すれば良い．したがって、統計的手法のように統計情報を得るために用いるコーパスの選択基準やコーパス量の問題は生じない．また、統計的手法では選択したコーパスでの確率で規則に重み付けを行っている．これに対し、本手法では順次入力される文によってルールの重み(尤度)は動的に変化していく．したがって、入力文の対象や言い回しなどが変化した場合にも対処することができる．本手法では，入力を品詞付単語列(一般的な形態素解析結果)とし，最初に係り受け単位区切処理を行う．次に，区切られた係り受け単位に対して，学習されたルールを用いて係り受け解析処理を行う．これらの処理で用いられるルールの学習は帰納的学習を用いて再帰的に行われる．このように，自動的に解析ルールを獲得することにより，さまざまな抽象度の解析ルールを獲得できる．獲得された解析ルールを用いて，入力文に対して最適な抽象度のルールを適用，つまり具体的なルールから順に適用することにより解析精度を向上させる．

本手法では，文法規則を人手により予め与えずに，品詞付単語列を構文解析結果に変換するルールを自動的に獲得する．これにより従来の文法規則では対処ができなかった会話文などの非文法的な文や省略の多い文にも柔軟に対処できる．また，一般的には構文解析が困難である，対象に依存した文の解析も，システムが動的に対象に適応することにより解決できることが多い．これは，このような限定された対象にのみ有効な解析ルールを学習により自動的に獲得することが可能だからである．さらには，構文解析例から解析ルールを獲得することが可能なので，適応能力が高く，さまざまな文章に対応できる．特に，本手法の学習アルゴリズムは言語に依存していないので，学習データさえあれば同一のアルゴリズム（帰納的学習）でどのような言語にも適用可能である．すなわち，多言語対応構文解析手法としての能力（言語非依存性）を有するものと考えられる．

本稿では，最初に本手法の概要について述べ，次に

作成した実験システムで行った評価実験の結果について述べる．さらに，評価実験の結果から本手法の言語非依存性と今後の研究課題について考察する．

2　処理過程

図1に本手法の概要を示す．各処理の詳細についてはすでに文献[6, 7]で述べているので，本稿では，本手法の概要のみを記述する．

最初に，品詞付単語列を入力する．英文は単語が分かち書きされているので，単語に品詞を付属させたものである．日本文の品詞付単語列は帰納的学習を用いた単語の認識手法[8]によって得られる単語列に品詞を付属させたものである．

次に，後述する学習処理部で獲得された係り受け単位区切ルールを用いて，入力品詞付単語列を係り受け関係を表すときの単位に区切る．表1, 2に入力の品詞付単語列と係り受け単位の例を示す．

さらに，係り受け解析部では学習処理部で獲得された係り受け解析ルールを用いて処理を行う．係り受け関係の例を表1, 2に示す．ここで，本稿で用いた各記号の意味を説明する．括弧で囲まれた部分が係り受け単位を表す．“/”の前が単語で、後ろのアルファベットは，その単語の品詞を表している．さらに，矢印はそれぞれ係り元単位から係り先単位に到達し，係り受け関係を表す．

上述の係り受け単位区切処理および係り受け解析処理においてルールを適用するときに複数のルールが適用可能な場合がある．適用するルールを決定するために以下の尤度評価関数を利用し，(A)式の値が最大となるルールを適用する．

$$\mathbf{CEF} = \mathbf{AF} + \alpha \times \mathbf{CF} - \beta \times \mathbf{EF} \cdots (\mathbf{A})$$

AF：出現頻度, **CF**：正適用度数, **EF**：誤適用度数

α, β：係数

この尤度評価関数が意味するのは，頻繁に出現し，精度が高く，誤りが少ないルールの尤度が高くなり，優先されて選択されるということである．

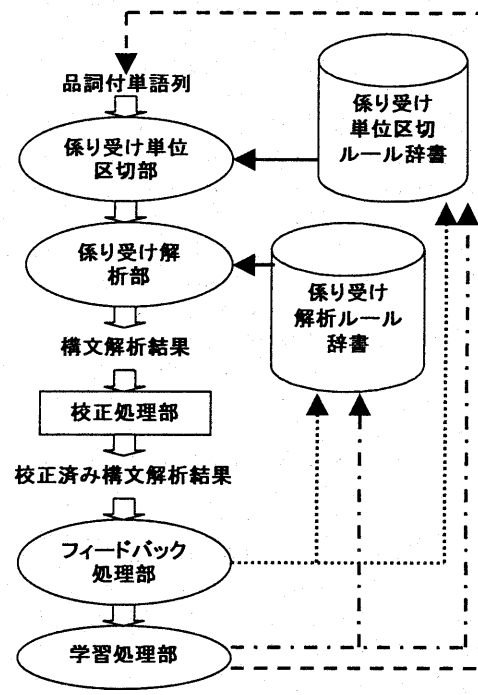

図1 処理過程

校正処理部では構文解析結果に誤りが含まれている場合に人手により校正が施され，校正済み構文解析結果が得られる．

フィードバック処理部では，係り受け単位区切部および係り受け解析部より出力された係り受け単位区切結果および構文解析結果と校正済み構文解析結果を比較することによりフィードバック処理を行う．正しい解析に使用されたルールの正適用度数に1を加算し，誤った解析に使用されたルールの誤適用度数に1を加算する．

学習処理部では，得られた校正済み構文解析結果か

表1 英文例

品詞付単語列	I/PRP want/VBP to/TO stay/VB here/RB one/CD more/JJR night/NN ./-PERIOD-
係り受け単位	(I/PRP) (want/VBP) (to/TO) (stay/VB) (here/RB) (one/CD more/JJR night/NN) (./-PERIOD-)
係り受け関係	(I/PRP) (want/VBP) (to/TO) (stay/VB) (here/RB) (one/CD more/JJR night/NN) (./-PERIOD-)

表2 日本文例

品詞付単語列	1/n 日/s 滞在/n を/p 延長/n し/v たい/s のです/x が/p . /Z
係り受け単位	(1/n 日/s) (滞在/n を/p) (延長/n し/v たい/s のです/x が/p . /Z)
係り受け関係	(1/n 日/s) (滞在/n を/p) (延長/n し/v たい/s のです/x が/p . /Z)

ら係り受け単位区切ルールおよび係り受け解析ルールを獲得する．さらに，獲得されたルールより帰納的学習を行い，新たな抽象化された文節区切ルールおよび係り受け解析ルールを抽出する．本手法の帰納的学習では字面の共通部分と差異部分を見つけ，差異部分を一般化(変数化)することによりルールを学習する．この操作は獲得されたルールに対して新たなルールが抽出されなくなるまで繰り返し行う．係り受け解析ルールは，ルールの組の共通度が一定の閾値以上の場合に獲得される．以下の式で，共通度は決定される．

$$\text{共通度}=\frac{\text{ルール中の同じ単語}}{\text{ルール中の全単語数(変数を含む)}}\times 100[\%]$$

ただし，ルール中の同じ単語と認定するのは，同じ順序で出現した場合とする．

また，抽象化されたルールから新たなルールを抽出するときは抽象度が一定の閾値以下の場合に行う．以下の式で，抽象度は決定される．

$$\text{抽象度}=\frac{\text{ルール中の変数の個数}}{\text{ルール中の全単語数(変数を含む)}}\times 100[\%]$$

共通度および抽象度による制限を加えないと過学習が進行し，解析時に使用できない係り受け解析ルールが多数獲得されてしまう．これにより処理効率の低下が起こる．したがって，本稿では係り受けルールを抽出するときに共通度および抽象度によってルールの獲得に制限を加えている．なお，共通度と抽象度の和を100[%]と定義した．これはルールのX[%]が共通部分であるとき，差異部分をすべて変数化した場合の抽象度が(1−X)[%]となることから決定した．

以上のような処理を繰り返し実行することにより，係り受け単位区切ルール辞書および係り受け解析ルール辞書が充実し，これらの辞書を用いて行う構文解析の精度が上昇する．

3 実験

本手法の性能，特に言語非依存性と入力に対する適応性を評価する実験を行った．実験では旅行用英会話文[9]の英文と日本文を用い、異なる言語を入力とすることで本手法の言語非依存性を調べる．なお、英文と日本文は対訳関係にある．英文と日本文を選択したのは，述語と目的語の位置が異なるなど文法的に大きく変わる言語に対して本手法の有効性を確認するためである．コーパスの内容は飛行機内、空港、およびホテルでの会話である．コーパスの内容を変化させることで入力に対する適応性能を調べる．

最初に実験手順を説明し，次に共通度，抽象度および尤度評価関数の係数を決定する予備実験について述べ，最後に評価実験結果を示す．

3.1 実験手順

初期状態では係り受け単位区切ルール辞書および係り受け解析ルール辞書は空である．これは，入力に対してどれだけの適応性能があるか評価するためである．以下の手順により実験を行った．

① 旅行用英会話文を品詞付単語列に変換し，システムに入力する．
② 入力された品詞付単語列について係り受け単位区切処理，係り受け解析処理を行い，構文解析結果を得る．
③ 構文解析結果に誤りが含まれている場合は人手により校正処理を行い，校正済み構文解析結果を得る．
④ 校正済み構文解析結果から係り受け単位区切ルールと係り受け解析ルールを獲得し，さらに帰納的学習を用いてルールを抽出する．
⑤ ④を新しいルールが抽出されなくなるまで再帰的に繰り返す．
⑥ 以上の操作を1文入力毎に繰り返し行う．

3.2 予備実験

まず，係り受けルールを獲得するときの最適な共通度あるいは抽象度を決定する実験を行った．実験では旅行用英会話文の英文および日本文500文ずつを用い，

表3 英文における共通度(抽象度)と解析率および係り受けルール数の関係

共通度(%)	抽象度(%)	正解析率(%)	誤解析率(%)	係り受けルール数
70	30	50.7	49.3	1,682
60	40	54.9	45.1	1,727
50	50	56.5	43.5	1,792
40	60	56.5	43.5	1,794
30	70	56.5	43.5	1,830

表4 日本文における共通度(抽象度)と解析率および係り受けルール数の関係

共通度(%)	抽象度(%)	正解析率(%)	誤解析率(%)	係り受けルール数
70	30	62.2	37.8	1,735
60	40	64.2	35.8	1,834
50	50	66.9	33.1	2,051
40	60	65.5	34.5	2,166
30	70	65.5	34.5	2,253

表 5 英文における係数と解析率の関係

α	β	正解析率(%)	誤解析率(%)
1	1	56.2	43.8
1	3	56.5	43.5
1	5	56.2	43.8
1	10	56.2	43.8
3	3	56.2	43.8
5	3	55.9	44.1
10	3	55.9	44.1

表 6 日本文における係数と解析率の関係

α	β	正解析率(%)	誤解析率(%)
1	1	65.5	34.5
1	3	66.9	33.1
1	5	66.2	33.8
1	10	66.2	33.8
3	3	65.5	34.5
5	3	64.9	35.1
10	3	64.9	35.1

共通度を 70%から 30%まで 10%毎に変化させた．ここで用いた尤度評価関数の係数は暫定的に，$\alpha=1$，$\beta=3$ とした．401 文から 500 文までの結果を表 3, 4 に示す．表 3, 4 より，英文では共通度が 50%以上で，日本文では共通度が 50%で，最も正解析率が高い．英文の場合，解析精度的に差異はないが，処理効率を考えると精度が同じなら獲得される解析ルールは少ないほうが良い．それゆえ，英文および日本文の係り受け解析ルールの獲得条件を，最も正解析率が高く，処理効率に優れると考えられる共通度 50%とし，以下の実験を行った．

次に尤度評価関数の係数の最適値を求める実験を行った．まず，α を 1 に固定し，β を 1 から 10 まで変化させて，共通度を決定するときに用いた旅行用英会話文の英文および日本文 500 文ずつについて実験を行った．実験結果から，英文，日本文の場合とも $\beta=3$ のときに最も正解析率が高かったので，β を 3 に固定し，α を 3 から 10 まで変化させ，さらに同様の実験を行った．実験結果を表 5，6 に示す．表 5，6 より，英文，日本文ともに $\alpha=1$，$\beta=3$ のときに最も正解析率が高くなった．

したがって，係り受け解析ルールの獲得するときの共通度を 50%，尤度評価関数の係数 α を 1，β を 3 として評価実験を行った．

3.3 評価実験結果

評価実験では旅行用英会話文 1,410 文の英文と日本文を用いた．実験に用いた英文の単語数は 9,622 で，1 文当たりの平均単語数は 6.8 単語，また係り受け数は 4,761 で，1 文当たりの平均係り受け数は 3.4 であった．日本語文の単語数は 10,322 で，1 文当たりの平均単語数は 7.3 単語，また係り受け数は 2,293 で，1 文当たりの平均係り受け数は 1.6 であった．図 2〜5 に評価実験結果を示す．図 2, 3 は英文および日本文の係り受け単位数 200 個毎，図 4, 5 は英文および日本文の係り受け関係 200 個毎の解析率の変化を表している．図中の直線はコーパスの内容が変化した位置を表している．また，評価実験終了後の各辞書中のルール数を表 7 に示す．

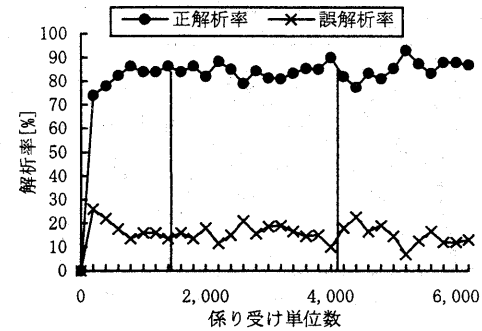

図 2 係り受け単位区切解析率の変化(英文)

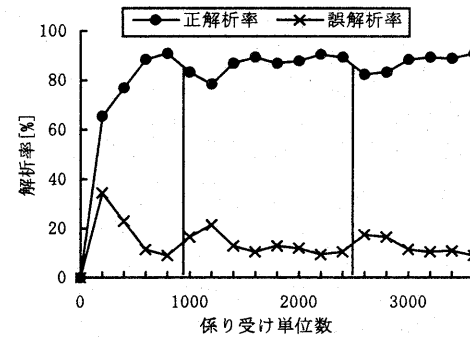

図 3 係り受け単位区切解析率の変化(日本文)

4 考察

予備実験の結果から共通度(抽象度)および尤度評価関数の係数が英文，日本文とも同じ値となった．しかしながら，一般的に言語によってこれらの値は変化するものと考えられる．それゆえ，これらの値についてはさらに詳細に検討する必要があると考える．

図 2, 3 の係り受け単位区切の解析率の変化を見ると，英文と日本文によって係り受け単位数の違いがあるものの同じような解析率の変化となっている．係り受け単位数が大きく異なるのは，英文の場合は 1 単語で係り受け単位となることが多く，それに対して，日本文は 2 単語以上で係り受け単位となることが多かったためである．また，コーパスの内容が変化すると一時的に正解析率が減少するが，再び正解析率が上昇した．これは学習が効果的に行われ，入力に適応することができたと考えられる．

図 4, 5 の係り受け解析率の変化を見ると，係り受け単位区切の解析率と同様に，係り受け数の違いはある

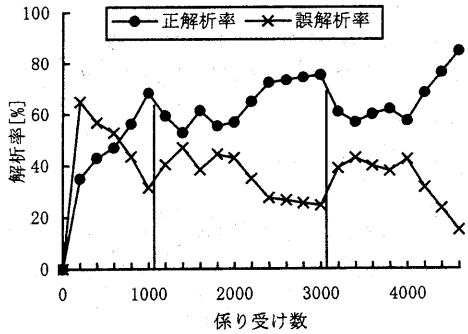

図4 英文における係り受け解析率の変化

表7 評価実験後の解析ルール数

ルールの種類	ルール数
係り受け単位区切ルール(英文)	1,685
係り受け解析ルール(英文)	5,631
係り受け単位区切ルール(日本文)	2,556
係り受け解析ルール(日本文)	6,713

が英文と日本文ともに同じような解析率の変化と考えられる．また，コーパスの内容が変化すると一時的に正解析率が減少するが，再び正解析率が上昇した．しかしながら，係り受け単位区切の解析率の上昇に比べると鈍い．これは適切な係り受け単位区切ルールを学習することよりも，適切な係り受け解析ルールを学習するのに多くの入力を必要としたと考えられる．ルールの形態からしても適切な係り受け解析ルールを獲得するためには多くの学習例が必要なのは明らかである．

また，図4，5から英文と日本文における係り受け正解析率は学習が進んだ状態ではどちらも 70%程度である．これは文法が大きく異なる言語である英語と日本語を解析して，どちらも同程度の精度で解析できたことを表している．ただし，どちらの言語においても個々の言語の従来手法と比較して精度が優れているとはいえない．しかしながら，ある特定言語にのみ適応したシステムはそれぞれの言語について作成しなければならない．本手法では基本的なシステムは種々の言語に対しても同じものが使用できるので非常に労力の軽減が可能となる．したがって，本手法には言語非依存性の能力が備わっていることを確認し，その有効性を見出すことができた．

5 おわりに

人手により与えられた構文解析規則を用いた構文解析手法と用例に基づく構文解析手法の問題点を解消するために，実例からの帰納的学習を用いた構文解析手法を以前に提案した．提案手法に基づくシステムを作成し，本手法の言語非依存性と入力に対する適応性能を評価する実験を旅行用英会話文を用いて行った．係り受け関係の正解析率は，学習が進んだ状態で英文および日本文において，ともに70%程度の評価実験結果を得た．これにより，本手法の言語非依存性を確認した．

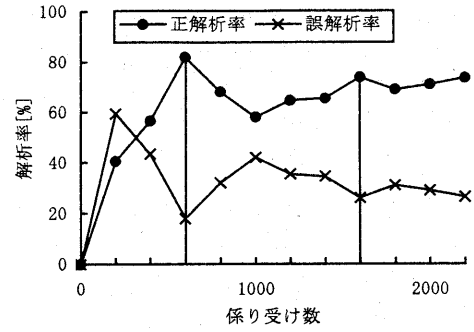

図5 日本文における係り受け解析率の変化

また，コーパスの内容を変化させて入力した時の正解析率の変化から本手法の入力に対する適応性能を確認した．したがって，本手法には言語非依存性の能力が備わっていることを確認し，その有効性を見出すことができた．

今後は，言語非依存性や適応性能といった本手法の利点を残して，さらに精度の向上を図るために解析ルールの学習および適用について検討する．

参考文献

[1] V. Sadler and Vendelmans: "Pilot implementation of a bilingual knowledge bank, Proceedings of the 13th COLING, pp. 449-451, 1990.

[2] C. Stanfill and D. Waltz: "Toward memory-based reasoning", Communications of the ACM, Vol. 29, No. 12, pp. 1213-1228, 1986.

[3] S. Sato and M. Nagao: "Toward memory-based translation", Proceedings of the 13th COLING, pp. 247-252, 1990.

[4] M. H. Al-Adhaileh and T. E. Kong: "A flexible example-based parser based on the SSTC", Proceeding of the ACL, pp. 687-693, 1998.

[5] 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, 1998.

[6] 正富欣之，荒木建治，栃内香次："実例からの帰納的学習を用いた構文解析手法"，信学技報，TL99-44, 2000.

[7] 正富欣之，荒木建治，栃内香次："模擬対話音声コーパスの書き起こし文における帰納的学習を用いた構文解析手法の性能評価"，情処研報，NL141-1, 2001.

[8] 荒木健治，栃内香次："帰納的学習による語の獲得および確実性を用いた語の認識"，電子通信学会論文誌 D-II, Vol. J75, No. 7, pp. 1213-1221, 1992.

[9] 地球の歩き方編集室：旅の会話集 2 米語/英語，ダイヤモンド社，1993.