社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－DBS－134　(II)　(79)
2004／7／14

検索エンジンを使った翻訳サポートシステムの構築

大鹿　広憲†　　佐藤　　学†　　安藤　　進††　　山名　早人†††

あらまし　本稿では、英作文を支援するためのサポートシステムの構築について述べる。ある文章を英作文したとき、作成した文章の中で文型的にどんな前置詞を使ったらいいのか、あるいはどの名詞にしたらいいかと迷うことがある。このような時、その文型が実際使われているかどうかをフレーズ指定として検索エンジンで検索すると、その文型の汎用性を調べることができる。以上の性質を利用し、本稿では、安藤進著の「翻訳に役立つ Google 活用テクニック」(丸善) で紹介されている手法を元に検索エンジンを利用した翻訳及び英作文の作業についてのサポートシステムを構築した。作成した英文の汎用性を調べるための様々な検索テクニックをシステム側で自動的に構築することによって、従来の Web 和英辞典よりも便利な Web における英作文支援のサービスを提供することに成功した。

キーワード　翻訳,Web サービス, 検索エンジン

A Translation Support System using Search Engines

Hironori OSHIKA†, Manabu SATOU†, Susumu ANDO††, and Hayato YAMANA†††

Abstract　This paper proposes a new Japanese-to-English-translation support system using search engines. The system uses Google to address some of the problems that non-native speakers of English will often encounter when trying to write English sentences. Especially, appropriate choices of English nouns and prepositions are a challenge for Japanese. To solve these problems, we focus on the techniques presented in the book titled "Using Google to Improve Your Translation Skills" written by Susumu Ando, published by Maruzen in 2003. The proposed system is implemented by automatically constructing optimal queries to check for proper English usage and has proved to be more convenient than using conventional Japanese-to-English dictionaries on the Web.

Key words　tarnslation, Web Service, Search Engine

1. はじめに

近年、英語の必要性はますます高まってきており、英作文をするといった機会が増えてきた。それに伴い、多くの和英辞典や翻訳システムが開発されてきている [10]。しかし、和英辞典だけでは十分な英作文ができず、また翻訳サイトや翻訳ソフトは、原文の構文と字句を反映させしまう直訳と呼ばれる訳文に近くなってしまう問題点がある。また、一つの日本語の名詞や動詞に対して複数の英単語が存在したり、前置詞も数多く存在するので、どれを使ったらいいか迷うことがある。

一方、インターネットの普及によって、検索エンジンの普及は高まり、誰もが検索エンジンを使って、情報を求める機会が増えてきている。しかし、検索エンジンは「情報検索」というのに利用するだけではなく、「表現検索」という形で利用することもできる。つまり、英作文の作業において検索エンジンを使うことによって、その文型の汎用性を調べることができる。例えば、「医療施設」を英語にする場合、「medical facility」「medical institution」の 2 通りが考えられる。どちらがより一般的に使われているかを調べるときに、これらの熟語を検索エンジン Google [1] でフレーズ検索を行うと、「medical facility」の方が検索ヒット件数が多いことがわかる。従って、「医療施設」の訳は

†, ††† 早稲田大学理工学研究科
Graduate School of Science and Engineering
†† 翻訳者
Translator

「medical facility」として使われるのが多いことがわかる。

しかし、このような作業においては、それぞれのフレーズを検索エンジンにより検索し、検索結果を見て比較するという手間がかかってしまったり、ワイルドカードを使用して検索を行ったときにそれぞれの検索結果を見ていくのが大変であるという問題点がある。

そこで、本稿では以上の問題点を解決するために検索エンジンを使った翻訳サポートシステムを構築した。本システムを構築することにより、各フレーズにおいて検索式を入力して調べるという手間が省くことができ、英作文作業において、前置詞や名詞の使い分けが分からないといった人達を支援することができると考えられる。検索エンジンは GoogleAPI を用いた [3]。

以下、2 節では英作文支援に関する関連研究について述べる。3 節では検索エンジンを使った翻訳の方法について述べる。4 節ではその検索テクニックを自動化するためのアルゴリズムについて述べる。5 節で評価を行い、6 節でまとめを行う。

2. 関連研究と研究目的

本節では、翻訳に関する関連研究について述べる。翻訳に関する関連研究は、主に 2 種類に大別できる。

- 機械翻訳
- 英作文支援

以下、それぞれについて述べ、それに伴う本研究の位置づけについて述べる。

2.1 機械翻訳の現状

機械翻訳は、定められた法則に基づいてデータベースを使って翻訳を行う。

しかし、日本語は複雑で、特定の単語や文について幾通りもの翻訳・解釈の仕方があるために、完全な翻訳が難しい。近年では、幾通りもの意味解析に対応した対訳コーパスの作成の研究が多くなされている [4] [5] [6]。

また、豊橋技術科学大の鈴木は、非対訳コーパスを用いた対訳語推定を情報検索システムと言語横断情報検索システムの検索結果から対訳語を推定し、100 件の名詞についての対訳語抽出実験を行っている [7]。翻訳に検索エンジンを用いている点は、本研究と非常に似ているが、名詞のみに着目しているため、汎用性に欠けるのが欠点である。

2.2 英文書作成支援

電気通信大学の高倉は、機械翻訳の質が低いことを指摘し、文書作成支援システム TransAid [8] を提案している。このシステムで、日本語の文章と、市販の機械翻訳システムによるその翻訳例を入力とする。次いで、翻訳システムの出力を訂正したり、洗練したりして、特定の目的に合った英語にするため、有用な英語の例文はインターネット・コーパスから抽出する。

インターネット・コーパスに検索エンジンを用いているところは本研究と似ているが、学会に関するページのみを集め、動詞名詞に限定した訂正を行っているため、前置詞を使った熟語やその他の品詞の分野に対して、汎用性がないのが欠点である。

他にも、数量表現の翻訳方法ついての支援 [9] や構文に関するスタイルのチェックなど、英作文支援についても既に数多くの研究がされているが、文章の分野が限定されていたり、法則を定義していてもツールとしては十分なシステムになっていないものが多い。

2.3 本研究の位置づけ

以上で紹介した手法は、翻訳の法則を自動化を実現するには難しいものであったり、実際に Web で提供されているものは少ないのが現状である。また、英作文のコーパスを構築する際にも、英語には様々なパターンがあるので、十分網羅できないといった欠点がある。

そこで、本研究では、翻訳の Web 上でのサービスを実現することを目的に、検索エンジンを用例データベースとして用いた名詞や動詞や助動詞などの訳語選択を自動的に提示できる翻訳サポートシステムを提案する。Web ページは、人手で作成されたものが多い。従って、検索エンジンを用例データベースにすることによって、42 億の Web ページを用例として参照できるほか、汎用性の高い文型を用例と共に検索ヒット件数で調べることができるという利点がある。

3. 検索エンジンを使った英作文の方法

本節では、文献 [2] を参考にし、検索エンジン Google を使った翻訳の方法について述べる。英語は、一つの英単語に対し、複数の日本語の意味が存在し、またその逆もある。「どの単語を使ったらいいのか」や、前置詞の使い方においても「この動詞における名詞にかかる前置詞はどれを使ったらいいのか」と迷うことがある。

このような場合、英作文した文章について部分的に気になっている文型の汎用性を検索エンジンを使うことにより参考にすることができる。

3.1 フレーズ検索

英語の単語の区切りにはスペースを用いているため、入力された文章をそのまま Google にクエリとして入力すると、全ての単語の AND 検索になってしまい、語順を保ったままの検索ができない。また、Google では頻繁に使われる言葉や文字 (「I」,「a」,「the」,「of」) はストップ語句として扱われるので、検索キーワードから自動的に除かれてしまう。

以上の現象を回避するために、テキストボックスに入力される文章に対して、フレーズ検索を行う。これにより、そのまま入力した文型がよく使われている表現なのかを調

査することができる。

3.2 前置詞の検討

例として、以下のような文章の英訳を行うとする。

現在製造されている自動車のほとんどはガソリンで走る。

まず、この文章を翻訳ソフト Excite[10] で英作文してみると、以下のようになる。

Most cars manufactured now run with gasoline.

以上の文章において、「ガソリンで走る」の前置詞は、「with」でいいのだろうかと気になる。そこで、「ガソリンで走る」という文章は、どのような前置詞が使われているのかを調べるために、

"run * gasoline"

と with の部分をワイルドカードに置き換えて検索すると、検索結果として表示された用例の中から、

on, by,

を使っているものがあった。Excite[10] で訳したときの前置詞「with」を含め、以上の 3 つを前置詞の候補とする。

次に、それぞれの前置詞を含めたフレーズ検索を行うと、結果は表 1 のようになった。

表 1 前置詞の検討

検索文字列	ヒット件数
"run on gasoline"	4,420
"run by gasoline"	131
"run with gasoline"	50

表 1 から、「on」を使った場合が一番ヒット件数が多いことがわかり、この文型がよく使われているということがわかる。逆に翻訳ソフトの「with」を使った用例は少ないことがわかり、あまりこの文型は使われていないということが判断できる。

従って、「ガソリンで走る」の英訳は「run on gasoline」とするのが適切であると考えられる。

3.3 ワイルドカードの複数指定

調べたい部分の英語をワイルドカードに置き換える方法について述べたが、ワイルドカードの個数を増やしながら検索する方法もある。例として、以下のような文章の英訳を行うとする。

空気は窒素と酸素から成り立つ

まず、この文章を Excite で訳すと、以下のようになる。

Air consists of nitrogen and oxygen.

「成り立つ」の部分の英語訳は「consists of」であることがわかる。この部分をワイルドカードに置き換えて、フレーズ検索してみる。(表 2)

表 2 のようにそれぞれ検索してみると、検索結果から「consist of」の他に「is made of」という熟語も使われていることがわかる。この 2 つの熟語についての汎用性を調べるために Google でフレーズ検索してみると、

表 2 ワイルドカード複数指定

検索文字列	ヒット件数
"Air * nitrogen and oxygen"	180
"Air * * nitrogen and oxygen"	193
"Air * * * nitrogen and oxygen"	251

表 3 フレーズ指定

検索文字列	ヒット件数
"Air consists of nitrogen and oxygen"	6
"Air is made of nitrogen and oxygen"	2

表 3 よりいずれも用例が存在しヒット件数にも大差がないことから、「成り立つ」の訳として、「consist of」「is made of」のどちらを使っても良いということがわかる。

3.4 和英辞典と Google を使った検討

名詞の検討は、前置詞や動詞とは違った検討方法がある。例として、以下のような文章の英訳を行うとする。

価格安定は重要な経済政策目標である

同じように、Excite[10] で試してみると、以下のようになる。

The price stabilization is an important economic policy goal.

「価格安定」という部分が気になったとする。「安定」という単語は複数あるので、どの単語が一番適切なのか迷うことがある。和英辞典で「安定」を調べると「stabilization」の他に「stability」「equilibrium」がある。従って、「価格安定」の訳は

price stabilization
price stability
price equilibrium

が候補として挙げられる。そこで、これらの英語をフレーズ指定して Google のヒット件数を調べると、表 4 のようになる。

表 4 「価格安定」の訳語

検索文字列	ヒット件数
"price stabilization"	14,700
"price stability"	156,000
"price equilibrium"	8,290

表 4 の結果から、「price stability」として使用するのが圧倒的にヒット件数が多いことがわかる。従って、「価格安定」の訳は「price stability」として使うのが適切であると考えられる。

また、和英辞典と Google を使った検討は動詞においても行うことができる。

3.5 形容詞の検討

形容詞の検討もこれまでに挙げた方法を応用して検討することができる。例として、以下のような文章を英訳を行

うとする。

レーザー・ポインターにおける最も重大な危険は、目に対する一時的影響に起因する事故である。

Exciteで試してみると、以下のようになる。

Most serious risk of the ability to set to a laser pointer is an accident resulting from temporary influence to an eye.

「重大な」という部分の訳について検討する。形容詞が来る位置にワイルドカードに置き換えて「most * risk」でフレーズ検索を行うと、検索結果から「serious」の他にも「important」が使われていることがわかる。それぞれのパターンにおいて、再度フレーズ検索を行うと、

表5 「重大な」の訳

検索文字列	
”Most important risk”	21,700
”Most serious risk”	4,310

表5より「重大な危険」の「重大な」の訳には、「important」も使えることがわかる。

「危険」に係る形容詞は、「重大な」以外のパターンは少ないので、「important」という形容詞が係っているものが検索結果として出てきて使えるということができた。しかし、実際は名詞に係る形容詞がたくさんありすぎて、検索結果から他のパターンを探せないといった場合がある。例えば、一時的影響は「temporary influence」となっているが、この形容詞の部分をワイルドカードに置き換えて、「* influence」としてフレーズ検索を行うと「social influence」や「bad influence」などが出てきてしまい、これだけでは「一時的な」の他の単語の候補を調べることはできない。この場合は、名詞の検討のときにも行った、和英辞典で調べてからフレーズ検索を行う。「一時的な」は和英辞典を調べると「temporary」「transient」「transitory」が出てくる。

表6 形容詞の検討

検索文字列	ヒット件数
”temporary influence”	675
”transient influence”	234
”transitory influence”	127

表6から、ヒット件数に余り差が見られないことが分かる。従って、どの単語も「影響」に係る「一時的な」として使用することができる。

4. 検索エンジンを使った翻訳サポートシステムの構築

3.でGoogleを使った翻訳の方法について述べたが、いくつかの問題点が存在する。

（1） フレーズ検索を行う際に「”」(ダブルコーテーション) を付加するのが面倒である。

（2） ワイルドカードを用いる際、品詞ができないと、関係のないものが検索結果に紛れてしまう。

（3） それぞれの英語をフレーズ検索して検索結果ウィンドウを切り替えながら、検索ヒット件数を比較するのは効率が悪い。

問題点1を解決するためには、入力された文に対してあらかじめフレーズ検索を行うように設定することによって解決できる。問題点2については、品詞分解を行うことで解決することができる。問題点3についてはそれぞれの結果を整理して提示する形にすれば、便利なシステムとして提供できると考えられる。従って、本稿ではGoogleを使った翻訳サポートシステムを構築した。

本システムは、検索データベースとしてGoogleAPIを用いた。また、文型をチェックする必要があるので、品詞の特定にはEric BrillのMontyTagger [11] を使用した。更に最適な名詞の提示には、和英辞典データベースにアクセスする必要があるので、Jim BreenのEDICT [12] を使用した。以下、Googleを使った翻訳サポートシステムの構成について述べる。

4.1 前置詞の検討の自動化

3.2では、Googleを使った前置詞の検討について述べた。この一連の作業を自動化する方法について述べる。

再度、3.2の例を挙げながら説明する。まず、ユーザはテキストボックスに気になる部分の英語文 (run with gasoline) を入力してもらう。ここで、文章全体を入力してしまうと、文章全体のフレーズ検索になってしまい、定型文でない限り検索結果が0件になる可能性が高い。従って、文章の一部分について入力をしてもらう。

そして、その中で検討したい部分においてドラッグで囲んでもらい、「送信」ボタンを押す。3.2の例の場合、前置詞を検討したいので「with」をドラッグで囲むことになる。(図1)

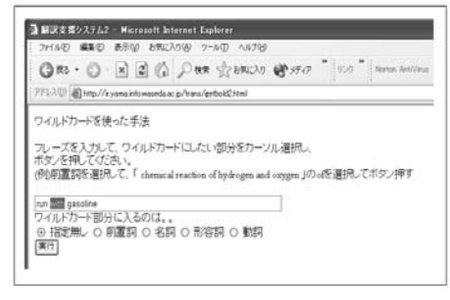

図1 システム画面

「送信」ボタンが押されたときのシステムの概要を図2

に示す。

- 入力された英文
- 選択した部分

の両方をシステム側に POST し、システム側は、選択した部分「with」をワイルドカードに置き換えて、Google でフレーズ検索を行う。このとき、「with」は品詞分解データベース MontyTagger から前置詞と判断する。そして、検索結果の snippet 部分の中で、＜b＞～＜/b＞タグで囲まれているところがクエリの部分なので、ワイルドカードに相当した部分の品詞を判定する。品詞分解データベース MontyTagger により、前置詞のものだけの用例を取り出すと「on」、「by」が使われているのがわかる。

システム側で、今度はクエリを「run with gasoline」と「run on gasoline」「run by gasoline」として GoogleAPI を用いて検索する。GoogleAPI による検索では一度に 10 件の検索結果しか得ることができないため、1 件目～ 10 件目、11 件目～ 20 件目とそれぞれ並行して GoogleAPI にクエリを送り並列で検索を行う。その後ヒット件数を表示する表を提示し、ユーザにどちらの前置詞がよく使われているかの提示を行う。

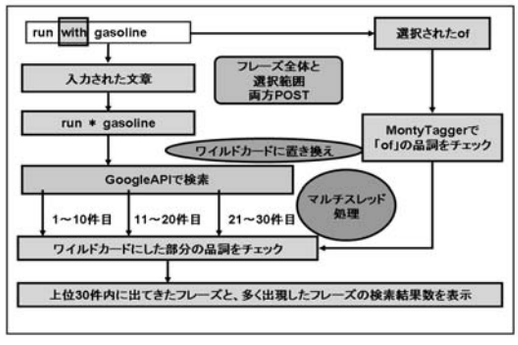

図 2　前置詞を検討するシステムの概要

前置詞の検討を行う場合のシステムの結果を図 3 に示す。

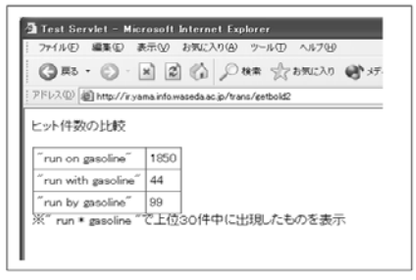

図 3　システム検索結果の画面

図 3 のように提示された情報を元にユーザは、最適な前置詞を判断することができる。

4.2　ワイルドカードを複数指定する方法

3.3 を実現する方法は、基本原理はワイルドカードを使う方法と同様なので、4.1 に部分におけるワイルドカードの指定が 1 個の時、2 個の時、3 個の時と置き換えるようにする。3.3 は、ワイルドカードの 1 個目が動詞であるという判定を行うことによって調べることができる。

4.3　和英辞典と Google を使った検討の自動化

3.4 で名詞の検討について述べた。本項では、3.4 の一連の流れを自動化する方法について述べる。

名詞の検討を行う場合、まずユーザには以下の 2 つの情報を入力してもらう

- とりあえず作成した「価格安定」という英語
- 調べたい英語の部分の日本語訳 (3.4 の例だと「安定」)

ここで、「安定」という部分においては、日本語と英語の両方を入力してもらうことになる。

「安定」という日本語を入力してもらうのは、和英辞典で類語を探すためである。日本語を入力した部分が和英辞典の辞書引きのためのクエリとなる。英語でも「安定」という単語を入力してもらっているが、英語から日本語に直す場合は何通りもの訳が考えられてしまう。また、入力してもらった英語に対してシソーラスから類語を提示する場合、和英辞典で辞書を引く場合と結果がかなり異なってくる。今回の場合は日本語から英語に直す作業においての支援システム作成なので、日本語における類語の提示が良いと考え、最適な名詞を探す場合においては、名詞の部分についての日本語訳を入力してもらうようにした。

英語で「価格安定」と入力してもらうのは、「安定」という単語がどういうときに使いたいのかの情報を得るためである。上記の場合は、「安定」という単語を「価格」と組み合わせて使いたい場合、すなわち安定という単語をどういうときに使いたいのかという情報を得るために、「価格安定」という単語を入力してもらうようにする。

次にとりあえず作成した「安定」の英語の部分をドラッグを囲む。これは、文型のパターンを読みとるためで、「price」の後に「安定」という単語が来るようなものを調べるために行う。入力形態と実行結果をそれぞれ図 4、図 5 に示す。

図 4　名詞の検討におけるシステム画面

従って、システムの流れは以下のようになる。(図 6)「安

定」という日本語を EDICT で辞書検索を行い、単語の候補の情報を取得する。次にその単語それぞれに対して、価格という単語と組み合わせたフレーズ検索を行う。そして、用例と共に検索結果数の提示を行い、最適な名詞を判断してもらう。

図 5　名詞の検討の実行結果

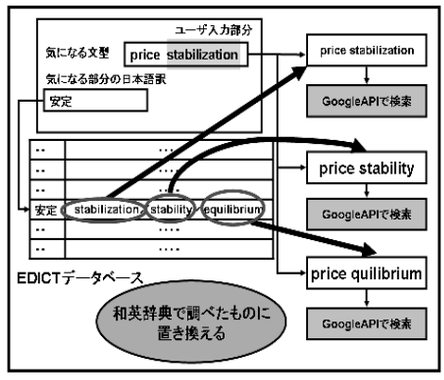

図 6　名詞検討するシステムの概要

4.4　形容詞の検討の自動化

形容詞の場合においても、前置詞や前置詞の検討のアルゴリズムとほぼ同様である。ワイルドカードにする部分は形容詞になるので、品詞の判断を形容詞にするという点で異なる。上記 2 つの機能を使い分けられるようにした。

5.　予 備 評 価

Web 上のサービスとしての提供なので、評価には機能、コスト、使いやすさ、応答時間などといった点が考えられるが、本稿では予備評価として機能的に使えるかどうかを評価した。

5.1　評 価 方 法

Web 上で使える翻訳システムの中で代表的なものである Excite の翻訳との比較を行った。任意の英作文集 [13] から選んだ日本語文と英語文を正解データとし、日本語文に対して Excite で英訳化を行い、実際にユーザが本システムを使って修正を施す。修正した部分が英作文集の英語文と内容がほぼ一致していれば、「**正解**」とし、修正を行っても正解データと著しく異なった場合は「**不正解**」とした。様々な英語表現があるので、本稿では「修正した部分」においてのみ正誤判定を行うとする。以下に例を示す。

英作文集による日本語訳と英語訳
日本語訳：
大学生は卒業までに最低 4 年間は授業に出席しなくてはいけない。
英語訳：
College students must attend classes for at least four years before graduating.

まず、日本語の文章を Excite の翻訳で英語翻訳を行う。

Excite による英語翻訳：
A college student has to attend a lesson for at least four years by graduation.

ユーザが「卒業までの 4 年間」の部分「four years by graduation」について、本システムで前置詞の検討を行った。すると、「four years before graduation」とする方が良いとわかり、修正した。英作文集の英語訳にも before を使っているので、前置詞の検討がうまくいったということになり本システムの有用性が示され、「**正解**」とする。

以上の方法で、英作文集から選んだ 142 文に対して評価実験を行った。

5.2　評価結果と考察

評価実験の結果を表 7 に示す。

表 7　評 価 実 験

検討した品詞	例文数	正解数	精度 (%)
前置詞	34	26	76.5
動詞	48	32	66.7
名詞	56	50	89.2
形容詞	56	36	64.3

前置詞の検討においては、高い精度を出すことに成功した。ワイルドカードに置き換えて検索することによって、的確な前置詞を見つけることに成功した。

名詞の検討においても高い精度を出すことに成功した。和英辞典と Google を組み合わせた検討は非常に有用性が高いことがわかった。

しかし、動詞と形容詞の検討においては精度が低かった。本システムでは検討する検索結果の対象を上位 30 件に絞ったが、名詞に係る形容詞は多種に渡るので、検索結果の上位 30 件では的確な形容詞を見つけられないことが多かった。

GoogleAPI は 1 日 1000 回という使用制限と、1 回のアクセスで最大 10 件の検索結果を返すという制限があるため、検討する検索結果数を増やすと本システムが 1 日に使える回数が減ってしまうという問題が起きてしまう。以上のことから、GoogleAPI の機能的な問題も考えられる。また、動詞の検討におけるワイルドカードを複数指定する方

法でも検索の範囲が広すぎてしまうため修正することができなかった例文が多かった。

修正が失敗した要因として、英作文集[13]が単語集のものなので、意図的に汎用性の低い単語を使った英文を掲載していたためであると考えられる。

6. おわりに

本稿では、検索エンジン Google を利用した翻訳及び英作文の作業についてのサポートシステムを構築した。従来の研究では、実用化されているものが少なかっただけに、今回の実装は大いに意義があるものと考えられる。本システムを構築することによって、英作文の作業においてどのような名詞あるいは動詞にしたらよいか、この動詞にはどのような前置詞を付加すればよいかなどの疑問点を解決することに成功した。

しかし、課題は多く存在する。主な今後の課題として、

- 応答時間の減少
- 複数形や 3 人称単数の対応
- 分野別の用例の提示

などが挙げられる。また、このほかにもさらに多くの機能を追加すれば、より使いやすい Web 上のサービスとして提供することができると考えられる。

なお本システムは、`http://ir.yama.info.waseda.ac.jp/trans/` にて公開中である。

参考文献

[1] Google
http://www.google.com

[2] 安藤進著, ”翻訳に役立つ Google 活用テクニック”, 丸善, ISBN4-621-07294-3 (2003.10)

[3] 山名早人監訳, 田中裕子訳,Tara Calishain & Rael Dornfest 著：”Google Hacks”, オライリー・ジャパン,ISBN4-87311-136-6 (2003.8)

[4] 道祖尾太祐, 村上仁一, 徳久雅人, 池原悟:”日英対訳パターンの自動抽出に向けて”, 情処研報,NL-153-15,pp.113-124,(2003)

[5] 金出地真人, 徳久雅人, 村上仁一, 池原悟:”結合価文法による動詞と名詞の訳語選択能力の評価”, 情処研報,NL-153-16,pp.119-124,(2003)

[6] 荒牧英治, 黒橋禎夫, 佐藤理史, 渡辺日出雄:”用例ベース翻訳のためのパラレルコーパスからの対訳対発見”, 情処研報,NL-144-4,(2001)

[7] 鈴木健二, 梅村恭司:”情報検索システムを利用した日英対訳推定”, 情処研報, NL-151-1,pp.1-6,(2002)

[8] 高倉佐和, 古郡廷治:”TransAid-英文書作成支援システム-”, 情処研報, NL-150-2,pp.7-14,(2002)

[9] 延原由高, 池原悟, 村上仁一:”接頭・接尾辞を含む数量表現の翻訳方法”, 情処研報,NL-142-11,pp.75-82,(2001)

[10] EXCITE 翻訳
http://www.excite.co.jp/world/

[11] MontyTagger Project
http://web.media.mit.edu/ hugo/montytagger/

[12] EDICT Project
http://www.csse.monash.edu.au/jwb/edict.html

[13] 松本茂監修:”速読速聴・英単語”, 増進会出版社, ISBN4-939149-10-2 (2002)