非線形光学材料情報における意味関係の抽出と情報の構造化

○宇陀則彦[†]
石川雅弘[§]
山本毅雄[†]
藤原 譲[§]

The Extraction of Semantic Relationships and Structuralization of information for Nonlinear Optical Material Design

Norihiko Uda[†]
Masahiro Ishikawa[§]
Takeo Yamamoto[†]
Yuzuru Fujiwara[§]

Abstract

Advanced utilization of information is required in many kinds of fields. Semantic processing is a serious problem to be solved for realization of thinking support systems in material design. This paper describes extraction of semantic relationships and structuraliztion of information. Semantic relationships can be extracted from full-texts by unification-based formalism with feature sturctures. This method covers syntactic analysis and semantic analysis integratedly. Extracted semantic relationships are structuralized by self-organizer in Information-Base Systems with Self Organizing Receptor Interconnections.

1　はじめに

近年、記憶媒体の大容量化にともない、フルテキストを代表とする原情報も大量に格納されるようになってきた。現在、この大量のフルテキストをいかに有効に活用するかが大きな課題となっている。材料開発などの専門分野においても、ハンドブック、論文集などから知識を獲得し、材料開発の研究に活用したいという要求が高まっており、検索、演繹推論だけではなく、類推、帰納推論、仮説推論などの高度な機能をもったシステムが必要とされている。

本研究では、研究開発を支援するための自己組織型情報ベースシステムの開発を行なっている。自己組織型情報ベースは、フルテキストなどの原情報から抽出した意味関係情報を自己組織的に構造化し、構造化された情報に対して、検索、ブラウジング、ナビゲーション、類推、帰納推論、仮説推論などを行なうシステムである。

本稿は、非線形光学材料情報を例にとり、自己組織型情報ベースシステムにおけるフルテキストを対象とした情報から意味関係の抽出と情報構造化について述べる。近年、特定の分野のテキストから特定の情報を抽出するために、テンプレート照合による手法が研究されているが [1][2]、本研究における意味関係の抽出は、日本語句構造文法 [3][4][5] を参考にして単一化による文法の形式化の手法を用いてテンプレート照合により行なう。

[†]図書館情報大学
[†]University of Library and Information Science
[§]筑波大学 電子・情報工学系
[§]Institute of Information Science and Elecrtonics, University of Tsukuba

2　単一化による形式化

単一化による文法の形式化は、変型文法における変型に代えて構造の共有という考えを活用している [5]。基本操作は単一化であり、単一化は構造の共有を実現する機構といえる。

伝統的な文法では、語や句の統語的範疇を表すのは、N、NP などの単一の記号であり、意味情報は別に記述されていた。それに対して、単一化に基づく形式化は、多次元構造を持つ素性構造を採用し、各語の統語情報だけでなく、意味情報、語や句の間の依存関係も同じレベルで記述できる。これにより、統語処理、意味処理、文脈処理を統合した枠組で処理できる。

素性構造は、素性名と素性値の対の集合である。

意味表現の言語

情報抽出においてテキストから最終的に抽出するものは意味表現である。意味表現の形式言語としてのシンタックスを示す。この言語では PST （部分項）とよばれる、属性と属性値の対の集合形式のデータ構造を採用している。

1. **ストリング** (*strings*) ··· 任意の文字列。
2. **役割ラベル** (*role labels*) ··· **rel**、**sbj**、**obj**、**index**、**restriction**、**event**...。
3. **情報子** (*infons*) ··· 属性名が役割ラベル、属性値がストリングもしくは意味式であるような PST。

   例：{**rel**/“示す”, **sbj**/“光ファイバ”, **obj**/“光カー効果”}
4. **意味ラベル** (*semantic labels*) ··· **soa**、**neg**、**index**、**imperative**、**probably** ...。
5. **意味式** (*semantic formulae*) ··· 次のように定義される要素が 2 つのリストである：
   (a) 第一要素が意味ラベルで、第二要素が情報子であるようなリストは意味式である。
   (b) 第一要素が意味ラベルで、第二要素が意味式であるようなリストは意味式である。

   例：[**neg**, [**soa**, {**rel**/“示す”, **sbj**/“光ファイバ”, **obj**/“光カー効果”}]]

ストリングはある概念を表すための用語を指示する。情報子は意味的情報の最小単位であり、どのような意味単位であるかと、その中で各個体が果たしている役割を表現する。意味ラベルは情報子もしくは意味式が表す情報がどのような状態 (肯定・否定、様相など) であるかを示し、意味式が意味表現としての基本単位となる。

意味ラベルによる情報子の構造規定

情報子は PST であり、一般的にはどのような属性をとるか (含むか) は自由である。しかしそれが表す意味に従った処理を施すには、それがどのような属性を持っているか明らかである方が良い。そこで各意味ラベル毎に意味式の第二要素の構造を規定することにする。これらは、[5]、[6] を参考にした。

soa 対象間の関係 (state of affair, 事態) を表現する情報子。
属性名として必ず **rel** を含み、その値で対象間に成り立つ関係の種類を示す。 その他の属性はその関係が成り立つ対象を示すものであり、その名前、数とも関係により異なる。

neg 意味式。
その否定を意味する。

index 属性名として **index**、**restriction**(のみ) を持ち、 **index** 属性値として対象を、**restriction** 属性値としては **index** 属性値を含んだ意味式をとる。
index 属性値で示される対象が、**restriction** 属性値の意味式を満たすものであることを示す。 次の例で示すように、この形の意味式は **index** 属性値が変数である場合にこそ大きな意味を持つ。

例： $[\mathbf{index}, \{\mathbf{index}/X, \mathbf{restriction}/[\mathbf{soa}, \{\mathbf{rel}/\text{“示す”}, \mathbf{sbj}/X, \mathbf{obj}/\text{“光カー効果”}\}]\}]$

3　意味関係の抽出

3.1　情報抽出プログラムの概要

情報抽出プログラムは、形態素解析の出力である形態素と品詞番号の対のリストを入力とし、一文ごとに形態素と品詞番号の対から与えられる素性構造に基づいて文全体の素性構造を構築していく。ただし、“完全なパーザ”では一つの文は最終的に一つの素性構造にまとまると考えられるが、ここで考えるプログラムでは、必ずしも一つの素性構造にまとまるとは言えない。各構成素が群化するのに充分な素性構造が与えられない場合、統合されることなく残ってしまう形態素があるからである。
これは欠点ではなく、ノイズのある文の処理や、本研究で対象としている「光学材料ハンドブック」のように特殊記号などのパースしにくい形態素を多く含む文でも、そのような形態素を (情報抽出の観点からみて関係ないならば) 無視するなどして、処理をすすめることができるということであり、システムに頑健性を与えるものである。

3.2　対象とする構文

情報抽出の対象として特定する構文を以下に挙げる。

コピュラ文

「A は B だ」の形の文をコピュラ文と呼ぶ。
コピュラ文はさらに (少なくとも) 記述文と同定文に分かれるとされ、記述文「A は B だ」における B は必ず属性であり、対象 A が属性 B を持つことを表す。ここで対象 A は個体であってもそれ自身属性であってもよい [7]。
例：

- 電気光学効果$_A$ は、電界により媒質中の光速が光の振動方向により異なる 現象$_B$ である。

一方同定文は主にある属性を満たす個体を同定するものであり、次の四つの形に分類される。

1. α の R は v だ。 $\cdots$ 半導体レーザ$_{\alpha}$ のもう一つの基本的な 特性$_{R}$ は 発振スペクトル$_{v}$ である。
2. α は R は v だ。 $\cdots$
3. v が α の R だ。 $\cdots$
4. α は v が R だ。 $\cdots$

これらはみな「α における R は v である。」ということを表す。

属性および性質

同定文「α の R は v だ」に含まれている情報は、上述した“R は v である”だけではない。同定文に限らず、「α の R」の文型が現れる時、これは“α は R という属性を持つ”ということを表していることがある。
次の文は、“フェリ磁性ガーネット結晶のベルデ定数”についての情報の他に、“フェリ磁性ガーネット結晶は、ベルデ定数という属性を持つ”ということも同時に表している。

- フェリ磁性ガーネット結晶 の ベルデ定数は、反磁性体に比べ二桁以上大きく高感度である。

ここで得られる情報は、用語の taxonomy において各用語が持つ属性、性質として格納される。用語の taxonomy とは用語の階層関係を束構造で表現したものである。

同値関係

文中に現れる、括弧(“(”と“)”)で囲まれた表記から、同値関係(異表記)を抽出することが出来る。たとえば、

- デポジション (堆積)、熱拡散、イオン交換、イオン注入、 エピキャシタル成長などが代表的なものであるが、新しい技術が年々開発されているのが現状である。

からは“デポジション”と“堆積”が同値関係にあることが分かる。この判断は一般的には正しくないが、括弧に先行する語(句)と括弧内の語(句)の品詞を比べることにより(ここでは共に名詞である)行なえる。

上下関係

次の二つの文

- 磁気光学効果$_1$ には ファラデー効果(1次磁気光学効果)$_2$ と
  フォークト効果(2次磁気光学効果)$_2$ がある。
- 素子$_1$ としては フォトダイオード$_2$、 フォトトランジスタ$_2$、 太陽電池$_2$ などがある。

からは、下線1の下位概念として下線2があるという情報を抽出できる。第一文の表す関係は造語規則(と同値関係)の利用によっても分かるものであるが、第二文の表す関係は造語規則からは分からない。
そこで上下関係の抽出のためには次の四つの構文を特定する。

1. A には B[と C]* がある。

2. A にはB[、C]* がある。

3. A としてはB[と C]* がある。

4. A としてはB[、C]* がある。

ここで[X]* はX の0個以上の繰り返しを表している。
これらの情報は用語の taxonomy の構成に利用される。

因果関係

因果関係の表現としては、“原因”、“結果”などの語による直接的な表示もあるが、複文による表示が主である。ここでは次の構文に特定する。

1. “用言の終止形”＋“と”＋“用言の終止形”
   例：「電気光学効果を有する結晶に電界を印加すると屈折率が変化する。」

これら以外の情報に関する構文も、既存のシステムを変更することなく漸次追加していくことができる。(実際にシステムに与えるのは、そのための素性構造である。)

3.3　システムの構成

情報抽出の流れは、

1. 形態素解析 (各形態素ごとに品詞番号を与える)。
2. 形態素と品詞番号に基づき素性構造を与える。
3. 素性構造と各種原理に従って形態素 (を表す素性構造) を群化する (この処理の流れにより、陰に構文木が生成される)。
4. 意味表現を取り出す。

の順に進む。
形態素解析についてはIFS(ICOT Free Software) の形態素解析プログラムを利用するので、本研究で設計するのはそこから先の処理である。ここでは特に、品詞番号ごとのデフォルトの素性構造の設計と素性構造のグルーピングである郡化および、それらを踏まえたプログラムの設計について述べる。

3.3.1　素性の設定

表1に本システムの素性構造で使用する全ての素性名とそのとり得る値を示す。
素性は次のような階層を成している。

- 素性
  - head
    - pos
    - paux
    - kat
    - gr
    - pform
  - subcat
  - adjac
  - adjoin
  - sem
  - morph
  - slash

表 1: 素性一覧

素性名	素性値	役割
head	pos, paux, kat, gr, pform の各素性	主辞素性を示す
pos	n, v, p, adv, adn, conj	品詞 (part of speech) を表す
paux	pos の値によって異なる	品詞の補助情報を表す
kat	mizen1, mizen2, mizen3, mizen4, renyo1, renyo2, renyo3, renyo4, renyo5, shusi, rentai, katei, meirei	用言の活用形を表す
gr	sbj, obj	文法的役割を表す
pform	“は”,“が”,“を”,“の”, ...	後置詞 (助詞) を表す
subcat	素性構造のリスト	下位範疇化する構成素の素性構造を表す
adjac	素性構造	隣接して下位範疇化する構成素の素性構造を表す
adjoin	素性構造	前接して修飾する主辞の素性構造を表す
sem	意味表現	意味を表す
morph	正書情報	文字面を表す
slash	素性構造のリスト	スラッシュ素性を表す

表 2: paux 素性の pos 素性値ごとの値

pos 素性値	paux 素性値	役割
n	com	一般名詞 (common noun) であることを示す。
	pro	代名詞 (pronoun) であることを示す。
	for	形式名詞 (formal noun) であることを示す。
	num	数詞 (numeric) であることを示す。
	per	人名 (personal name) であることを示す。
	prp	人名以外の固有名詞 (proper noun) であることを示す。
	sah	サ変動詞性名詞であることを示す。
	na	「－な」の形がある名詞であることを示す。
	sahtsr	「－とする」の形があるサ変動詞性名詞であることを示す。
	ttr	「－と」「－たり」の形のみがある名詞であることを示す。
	post	人名について役職を示す名詞であることを示す。
	key	キーワードとして登録されている名詞であることを示す。
v	kd	「－の」の形がない形容動詞であることを示す。
	kdno	「－の」の形がある形容動詞であることを示す。
	kdnari	「－なり」の形がある形容動詞であることを示す。
	ky	イ段以外の音で終る形容詞であることを示す。
	kyi	イ段の音で終る形容詞であることを示す。
	kysouda	助動詞「そうだ」が接続する特殊形の形容詞であることを示す。
	dan5	五段活用の動詞であることを示す。
	u1dan	上一段活用の動詞であることを示す。
	s1dan	下一段活用の動詞であることを示す。
	ka	カ行変格活用の動詞であることを示す。
	sa	サ行変格活用の動詞であることを示す。
	kuru	「来る」が後接する動詞であることを示す。
	suru	「する」が後接する動詞であることを示す。
	zuru	「ずる」が後接する動詞であることを示す。
p	ない	
adv	ない	
adn	ない	
conj	beg	文頭にのみ出現する接続詞であることを示す。

表2に各 pos 素性値ごとの paux 素性値とその役割を示す。

また、slash 素性の値はその統語範疇 (が下位範疇化するべきであるがする相手がいなかったた

め) に足りない素性構造を表す。

表 3: デフォルトの素性構造の例

品詞番号	品詞名	素性構造
10	一般名詞	{head/{pos/n, paux/com}, morph/M, sem/M}
11	代名詞	{head/{pos/n, paux/pro}, morph/M, sem/M}
12	形式名詞	{head/{pos/n, paux/for}, morph/M, sem/M}
13	数詞	{head/{pos/n, paux/num}, morph/M, sem/M}
14	人名	{head/{pos/n, paux/per}, morph/M, sem/M}
15	人名以外の固有名詞	{head/{pos/n, paux/prp}, morph/M, sem/M}
16	サ変動詞性名詞	{head/{pos/n, paux/sah}, morph/M, sem/M}
17	形容動詞の語幹、 「－の」の形があるもの、 「正当(だ)」「親密(だ)」など	{head/{pos/v, paux/kdno}, morph/M, subcat/[{head/{pos/p, pform/PP}, sem/S}], sem/[soa, {rel/M, sbj/S}]} ; pform_ha_ga(PP)
18	形容動詞の語幹、 「－の」の形がないもの、 「明確(だ)」「古典的(だ)」など	{head/{pos/v, paux/kd}, morph/M, subcat/[{head/{pos/p, pform/PP}, sem/S}], sem/[soa, {rel/M, sbj/S}]} ; pform_ha_ga(PP)
19	名詞で「－な」の形があるもの	{head/{pos/n, paux/na}, morph/M, sem/M}
20	形容動詞の語幹、 「－なり」の形があるもの、 「愚か(なり)」など	{head/{pos/v, paux/kdnari}, morph/M, subcat/[{head/{pos/p, pform/PP}, sem/S}], sem/[soa, {rel/M, sbj/S}]} ; pform_ha_ga(PP)
21	サ変動詞性名詞、 「－とする」の形があるもの、 「生き生き(と)」「すべすべ(と)」など	{head/{pos/n, paux/sahtsr}, morph/M, sem/M}

3.3.2　情報抽出の例

例として次の文を用いる。

- 磁気光学効果にはファラデー効果がある。

これからは形態素解析により次のような形態素と品詞番号の対のリストを得る：
[[“磁気光学効果”|511], [“には”|452], [“ファラデー効果”|511], [“が”|420], [“あ”|48], [“る”|188], [“。”|2]]
さらにこれから素性構造が与えられ、それに従って群化が行なわれる。
以下の図でその過程を説明する。ただし形態素以外の **morph** 素性値は省略する。

まず図 1で、“磁気光学効果”と“には”が群化し、新しい素性構造を作る様子を示す。

これは“には”が **adjac** 素性値によって“磁気光学効果”を隣接下位範疇化することにより起こっている。主辞素性原理により、主辞の **head** 素性値はそのまま親の **head** 素性値になっている。また、意味素性原理により親の **sem** 素性値は主辞の **sem** 素性値と等しくなるが、主辞の **adjac** 素性値が娘と単一化することにより、娘の **sem** 素性値が親に伝わっている。

続けて入力として“ファラデー効果”が入ってきても、**subcat**、**adjac**、**adjoin** などの、群化を引き起こす素性値はないのでなにも起こらない。更に“が”が入ってくると、その **adjac** 素性値によって図 2のような群化が起こる。

これで二つの素性構造が出来た。ここに動詞の語幹“あ”が入ると、**subcat** 素性値に従って図 3のような群化が起こる。ただし、ここで先に“ファラデー効果が”が群化されないのは、なるべ

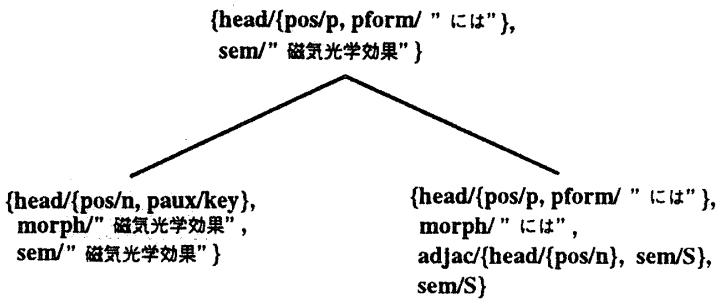

図 1: 「磁気光学効果　には」の群化

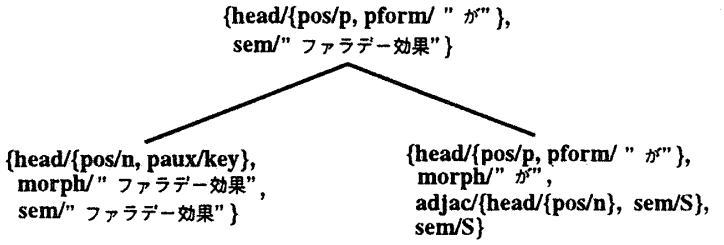

図 2: 「ファラデー効果　が」の群化

く形態的に近い場所の相手から群化することにしているためである。

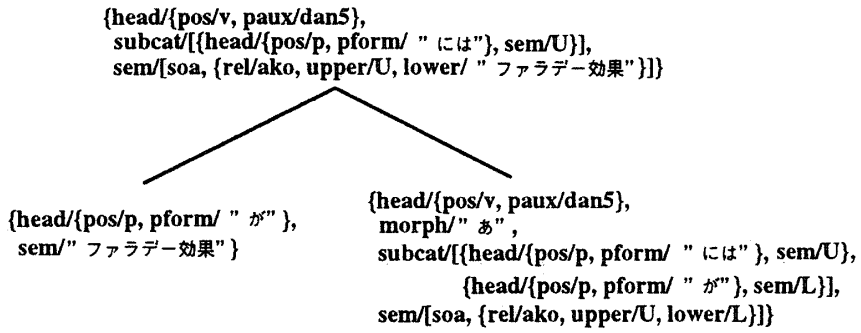

図 3: 「ファラデー効果が　あ」の群化

さらに subcat のもう一つの要素に従って図 4の群化が起こる。

最後に動詞の活用語尾“る”が入ることにより図 5の群化が起こる。

文末のマークである“。”が入ることにより、この文の処理が終了する。その時、具体値が代入された意味表現 (sem 素性値) があれば、それを抽出情報として出力する。ここでは [soa, {rel/ako, upper/“磁気光学効果”, lower/“ファラデー効果”}] が出力として取り出される。これは動詞の語幹“あ”の sem 素性値として与えられたテンプレートに、具体的な値が代入されたものである。

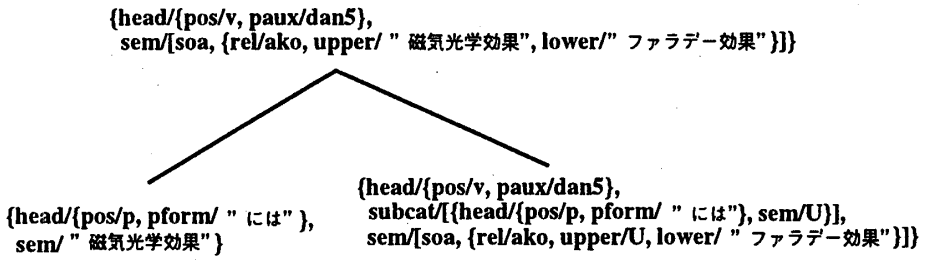

図 4:「磁気光学効果には　ファラデー効果があ」の群化

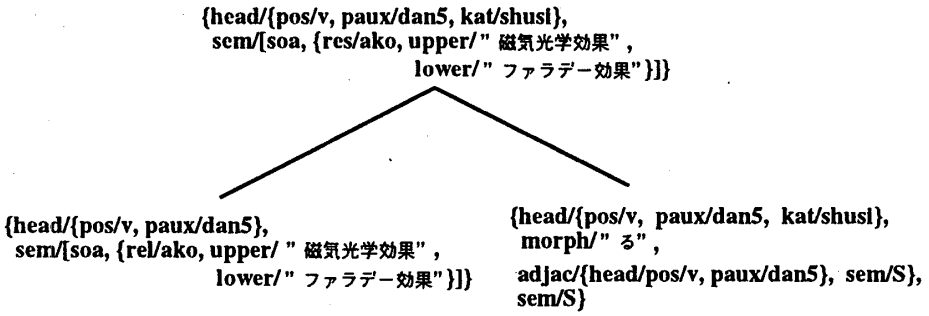

図 5:「磁気光学効果にはファラデー効果があ　る」の群化

このように、抽出したい情報に関して構文を特定するような素性構造を与えることにより、それがテンプレートの役割を果たし、情報抽出が行なえる。

4　考察

言語を処理する上で必ず直面する問題は、一つの意味に対して複数の表現があるという表現の多用性（同義性）、逆に一つの表現に対して複数の意味があるという意味の多用性（多義性）、さらに、表現が基本的には 1 次元という線形構造であるのに対して、意味は多次元であるというような表現と意味の構造の相違である。

これまで、これらの言語の問題に対して、目的に応じてさまざまな試みが行なわれてきた。検索用シソーラスにおいては、統制語を組織化して表現の多用性に対応しようとした。また、機械翻訳においては、辞書を組織化して各用語に意味素を付与することにより、意味の多用性に対応しようとした。これらは部分的には有効なのだが、言語全体を完全に処理するにはいたっていない。

自然言語の解析は、古典的、伝統的には、形態素解析、構文解析、意味解析、文脈解析というように段階的に別々に行なうのだが、各段階は不可分であることがわかり、最近では、同時に並行して行なわれるようになった。単一化に基づく形式化は、音韻情報をはじめ、形態素情報、統語情報、意味情報、依存情報を同じレベルで記述し、処理するうえで有効な手段である。

単一化に基づく形式化では、各用語、句、文に対して、形態素情報、統語情報、意味情報、依存情報を同格に表現した素性構造を付与し、各用語、句、文の範疇を多次元空間の点として表現できる。これにより、構文と意味を同レベルで処理することが可能になり、構文の処理が同時に意味構造の構築になる。また、素性構造を抽出したい情報に応じて用意することにより、特定の専門分野

に関わらず情報抽出が行なえる。

さらに、自己組織型情報ベースシステムでは、抽出した意味関係を利用して新しい構文パターンを抽出し、自己増殖的に意味関係と構文パターンを抽出できる [9]。

5　結論

本稿は、非線形光学材料情報を対象に、「単一化に基づく形式化」によって意味関係を抽出する手法について述べた。本手法は、単純なテンプレートによる手法に対して、構文的に誤った照合を防ぐことができる点が大きな特徴である。抽出された意味関係は、自己組織型情報ベースシステムにとりこまれることにより、類推、帰納推論、仮説推論のための構造情報として格納される。

参考文献

[1] Kitani, Tsuyoshi Eriguchi, Yoshio Hara, Masami: *Pattern Matching and Discourse Processing in Information Extraction from Japanese Text*, Journal of Artificial Intelligence Research, 1994, pp.89-110

[2] 江里口善生, 木谷強: 「富田一般化 LR パーザを用いた情報抽出」 情報処理学会, 自然言語処理研究報告, vol.102, No.2, 1994, pp9-16

[3] 三吉秀夫, 郡司隆男, 白井英俊, 橋田浩一, 原田安也: 「日本語の句構造文法 — JPSG」 コンピュータソフトウェア, vol.3, No.4, 1986, pp.39-45.

[4] Gunji, Takao: *Japanese phrase structure grammar*, D.Reidel Publishing Company, 1987.

[5] 郡司隆男: 「自然言語」 日本評論社,1994.

[6] 郡司隆男 (訳): 「HPSG 入門 制約に基づく統語論と意味論」 産業図書株式会社,1994. (Pollard, Carl & Sag, A., Ivan: *Information-based syntax and semantics, volume 1:Fundamentals*, Center for the Study of Language and Information, 1987.)

[7] 坂原: 「役割、ガ・ハ、ウナギ文」 認知科学の発展 Vol.3, 日本認知科学会 (編), 1990, pp29-66

[8] 安藤司文, 益谷真, 守屋秀洋: 「多様な自然文からの知識獲得」 情報処理学会, 自然言語処理研究報告, vol.98, No.4, 1993, pp25-32

[9] Sano, H. and Fujiwara, Y.: Syntactic and Semantic Structure Analysis of Article Titles in Analytical Chemistry. Journal of Information Science, 1993.

[10] 賴 静娟, 王 暁晶, 陳 漢雄, 藤原 譲: 「概念間の意味関係の自動抽出法とその応用例」 情報知識学会 第 1 回研究発表会 1993.

[11] 宇陀則彦, 藤原 譲: 「非線形光学材料を対象とした自己組織型情報ベースシステム」 情報知識学会 第 2 回研究発表会 1994.

[12] 賴 静娟, 王 暁晶, 藤原 譲: 「専門用語における階層関係および関連関係抽出法」 情報知識学会 第 2 回研究発表会 1994.