自然言語処理 80－5
（1990．11．22）

インドネシア語理解のための計算モデル

Hartono, 田中穂積
東京工業大学・工学部・情報工学科

概要

現在，自然言語の解析には，二つの流れがある．第一は，言語が生成し得るすべての構文構造に対して規則を用意するやり方である．第二は，辞書情報を豊富にし，規則の数を最小限に抑える方法である．前者は，膨大な数の規則を扱わねばならないことが多く，規則間に矛盾が生じたり，構築される体系が冗長になったり，開発・運用に大きな問題がある．後者は，HPSGに代表される理論を利用した方法であるが，普遍な規則を用いるため，範疇間の制約関係の記述が複雑になる．本研究では，語義辞書を利用した語義ベースパーザによる解析法を提案する．語義辞書には，曖昧性解消に必要な語彙共起，統語規則，意味及び文脈の制約情報が一括して記述される．この解析法に基づき，インドネシア語の具体的な解析アルゴリズムについて述べる．

A Computational Model for Indonesian Understanding

Hartono, Hozumi TANAKA
Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-Ku, Tokyo-152 Japan

Abstract

In this paper, an Indonesian analyzing system using a wordsense-based parser is described. The input sentence will be segmented first by considering the Indonesian keyword. The meaning of each segment is then examined and synthesized together by refering the informations, including the syntactical rule, described in each word-sense. By parsing only the word-senses those related to the input sentence, the unrelated futile searching problem, occured in rule-based approach which presupposes an enormous amount of rules, can be avoided. The syntactical rules are explicity applied so that, the intricate problem of describing the relationship among categories occured in the approach like the HPSG-based analyzing can also be solved.

1 はじめに

計算機による自然言語理解の研究とその実現に対する期待は，計算機技術の発達に支えられ，高まる一方である．計算機の誕生と共に計算機による自然言語理解への追求が始まったといわれている．しかし，計算機技術が目覚しい進歩を遂げているのに対し，計算機による自然言語理解の研究は，必ずしも期待に沿う成果を得ていないのが現状であろう．自然言語は，言語学的な立場のみならず，心理学的また哲学的な立場からの研究も進める必要があるという点からも，計算機による自然言語理解は簡単なものではないわけがうなずける．しかし，より高度な自然言語理解システムが実現できるよう，われわれは，これまで歩んできた路線に沿って改善だけでなく，観点を変えてより強力な道具を作成する必要がある．

これまでの計算機による自然言語理解は，使用されている言語理論から計算機上での処理の方法論まで英語を中心に考えてきた．本研究は，これまで計算機による言語処理に関わりの薄かったインドネシア語を取り挙げる．

計算機による自然言語理解の目的は，与えられた単語の列である言語表現から，その言語表現が持つ意味内容を抽出することにある．言語表現には，一般に単語の語義の曖昧性，構造的な曖昧性などが含まれている．言語理解のための計算モデルにおいては，いかにこれらの曖昧性を解消できるかが重要な問題であると同時に，これらの曖昧性解消に必要な情報をいかに与え，またそれを利用してどのような解析方法で行なうかも大切である．現在の自然言語の解析には，二つの流れがある．第一は，言語が生成し得るすべての構文構造に対して規則を用意するやり方である．第二は，辞書の情報を豊富にし，規則の数を最小限に抑える方法である．前者は，膨大な数の規則を扱わねばならないことが多く，記述される規則の間に矛盾が生じたり，構築される規則の体系が冗長になったり，開発・運用に大きな問題がある．後者は，HPSG [8] に代表される理論の枠組を利用した方法であるが，普遍な規則が使われているため，範疇間の制約関係の記述が複雑になる．

本研究では，これらの問題点を考慮して，語義ベースパーザと呼ばれる解析法を提案する．この解析法に基づき，インドネシア語の具体的な解析アルゴリズムについて述べる．解析では，まず，インドネシア語の特徴のある単語をキーワードに用い文を分割する．語義辞書の情報を手がかりに，各断片の意味を計算し，断片間の結合を行い，文の意味を抽出する．語義辞書には，曖昧性解消に必要な，語彙共起と統語規則と意味及び文脈の制約情報が，形態論と統語論と意味論との関係を統一的かつ明示的に表現できるように考慮して記述される．

2 インドネシア語の特徴

(1) Kemarin Adi memetik bunga di taman.
yesterday Adi pick flower in garden
(昨日アディは花園で花を摘んだ)

上記のインドネシア語の例文を見れば分かるように，インドネシア語の構文構造は，英語と似ている．しかし，インドネシア語の副詞的句が現れる場所は，英語と比べて比較的自由である．例えば，上の例文は，以下のように表現できる．

(2) *Di taman kemarin* Adi memetik bunga.
(3) Adi memetik bunga *di taman kemarin*.
(4) *Kemarin* Adi memetik bunga *di taman*.
(5) *Kemarin di taman* Adi memetik bunga.

インドネシア語の修飾語と被修飾語の語順は，英語・日本語と違い，被修飾語が修飾語よりも先に来る．例えば，「私の花」は"bunga saya"となる．また，接辞を有力な文法要素としていることもインドネシア語の大きな特徴である．基語が接辞を付加することによって基語の品詞および意味を変えることができる．以下では，基語 *dengar*(聞く)を基にして作り出される分出語の一部を示す．

接辞	分出語	意味
meN	men*dengar*	聞く
meN-i	men*dengar*i	注意して聞く
meNper-kan	memper*dengar*kan	聞かせる
di	di*dengar*	聞かれる
ke-an	ke*dengar*an	聞こえる
ter	ter*dengar*	ふと聞く
peN-an	pen*dengar*an	聴覚

インドネシア語は，英語のように動詞の変形によって時制を表現することがない．上に示した文中の"Kemarin" (yesterday) を取り除けば，残りの文が示すイベントは，過去に起きたものか，あるいは現在のものかを区別できない．時制を伝える手段として上記の文のように直接的に時間の副詞を入れるかまたはアスペクトマーカ[1]を利用する方法がある．文中にこれらが陽に表現されない場合には，文の時制は文脈を考慮する必要がある．

英語では，一般に動詞を手掛かりにして文を決定することができるが，インドネシア語の文では，必ずしも動詞が存在するとは限らない．例えば，

"Bunga itu$_{Np}$ di taman$_{Pp}$.(あの花は花園にある)",
"bunga itu$_{Np}$ cantik$_{Adj}$. (あの花はきれいだ)"

などがある．英語では，このような文は *be* 動詞を使って表現する．インドネシア語では，"Np Np"は，文となるか，または一つのNpともなり得る．これらの違いは，接尾辞-*nya*，指示詞でもある *itu*，または副詞などをうまく利用して区別できるようになっているのもインドネシア語の大きな特徴である．本研究では，それらの特徴を利用して文の解析を行う．

3 解析モデル

3.1 これまでの研究と問題点

自然言語の解析は，一般に，以下のようなプロセスがあるとされている．

- 文を構成する最小単位である単語の認定を行い，動詞などの活用変化形や複数名詞などの検査を行なう形態素解析と，

[1]例えば，"Adi *akan* memetik bunga.(アディは(これから)花を摘む)"の *akan* はアスペクトマーカの一つで英語の"will"の意味と同じ．

- 認定された単語列の文法的妥当性を検査し，それらの文法的役割を決定する統語解析と，
- 文の意味的妥当性及び文章の意味的整合性を検査し，文の意味内容を抽出する，意味解析と文脈解析

これまでの解析方法の一つに，上記の解析過程を逐次的に行う方法がある．これは，形態素解析の後に構文解析を行い，構文解析が終ってから意味解析を行うというやり方である．しかし，このように逐次的に解析を行うと，各過程が分離して処理できるという利点がある一方，構文解析の段階で構造的な曖昧性のために組合せ的な爆発を招き，効率よく解析を行えないという問題がある[2]．

一方，人間の自然言語理解のプロセスを考えてみると，人間は，上記に挙げられたプロセスを逐次的に処理しているということはなく，むしろ各プロセスが相互的に関連し合って曖昧性を解消していくというプロセスを行っていると思われる．また，統語的な曖昧性を解消するのに意味情報の利用が効果的であることを考えれば，これらのプロセスを独立的に行うのではなく，融合して解析した方が自然である．この方法によると，統語解析の段階で意味解析を導入するということから，上記に述べた統語解析による組合せ的な爆発の問題は極力的に押えることができるという利点もあって，多くの解析システムがこの方法を利用している[5, 6]．

上記の解析方法の多くは，規則駆動による古典的な統語解析方法を利用している[3]．しかし，この方法では，言語が生成し得るすべての構文構造をあらかじめ想定している．そのため，膨大な探索空間が必要で，また，膨大な数の規則を扱わなければならないことから，規則相互間に矛盾が生じたり，構築された体系が冗長になったりするなど[4]，開発とその運用に大きな問題がある．

一方，記述される統語規則を最小限に抑え，辞書の情報を豊富にする，HPSGに代表される言語理論の枠組を利用した解析方法がある．しかし，この方法によると，普遍的な規則が用いられているため，範疇間の制約関係の記述が複雑になり，実際のインプリメントにおいてかなりの工夫が要求される[9]．

また，逆に統語規則を利用せず，単語間の意味的結合性(概念依存構造)を頼りに直接意味抽出を行う研究が，Wilks[10],Rieger ら [7] によって行われている．例えば，Rieger らのワードエキスパートパーザの方法は，以下のような特徴を持つ．

- ワードエキスパートは，各自に必要な知識を独立に持つことで，統語規則を基にした解析法のように直接関係のない規則の間に不明確な相互作用と依存関係が存在するようなこともなくなり，記述にもモジュール性を持ち，明瞭である．
- 各ワードエキスパートに記述されている必要な情報量が大きいが，文に現れるワードエキスパートの数だけを解析に用いるため，実際に文の解析に必要な解析空間が小さい．

ワードエキスパートが行なう動作と，ワードエキスパート間の結合性でもって曖昧性解消を行なう，という基本原則であるが，単語間の複雑な結合性を網羅的に調べ出せるかどうかの問題が残る．この方法では，上に述べたような技術的な問題が残っているが，徹底的な概念の操作によって意味処理を行っている点は興味深い．

3.2 われわれの手法

ここでは，言語が生成し得るすべての構文構造に対して規則を用意する方法を用いたインドネシア語の解析の問題点を考えてみよう．インドネシア語は，第2章で述べたように，副詞句の出現は，随意的であり，かつ出現する場所も自由である．ここでまず，動詞Vが適用できる動詞パターンを"文 → Np,V,Np."とする．前置詞句 Prepp と副詞句 Advp とが随意的に付け加わることができると考えれば，動詞Vに対して少なくとも以下のような文法規則の記述が必要になるであろう．

文 → Np,V,Np.	文 → Np,V,Np,Advp,Prepp.
文 → Adv,Np,V,Np.	文 → Np,V,Np,Prepp,Advp.
文 → Np,V,Np,Advp.	文 → Advp,Np,V,Np,Prepp.
文 → Prepp,Np,V,Np.	文 → Prepp,Np,V,Np,Advp.
文 → Np,V,Np,Prepp.	文 → Advp,Prepp,Np,V,Np.
	文 → Prepp,Advp,Np,V,Np.

さらに，前置詞句が複数あったり，アスペクトなどその他の随意的な要素が加わればこれらの規則も組合せ的に増える．このように，インドネシア語が表現し得るすべての構文構造を用意するとなると，膨大な数の構文規則を記述する必要があり，前節で述べたような問題が生じる．

本研究では，インドネシア語の特徴と，上記の問題と，これまでの研究を踏まえて，以下の考え方を基にインドネシア語解析モデルを組立てる．

- 解析は，文を構成する単語と直接に関係のある語義だけを扱う．
- 統語規則について
  基本的な統語規則は必要であるが，規則を辿って解析するのではなく，統語規則は，あくまでも，曖昧性解消のための情報の一部として利用する．
- 曖昧性解消に必要な語彙共起，統語規則，意味，文脈の制約情報を明示的に一括して記述する．
- 情報は宣言的に記述する．
  副詞などの任意的なカテゴリが出現する場所を宣言的に記述することによって組合せ的に規則を列挙する必要がなくなるとか，優先性の記述が簡単にできるとかが利点として考えられる．

[2]なお，最近の研究では，この組合せ的な爆発を防ぐために，機械翻訳システムMuの文法 [1] では，すべての可能な統語解釈を出力せず曖昧のままにして置く方法を取っている．また，Maruyama[4]は，統語解析で得られるすべての可能な構造を一つにパックし，その構造に動的に制約を当てはめてみて，制約に違反したものを排除していくという方法を提案している．

[3]解析の戦略についての詳細は [2] を参考していただきたい．

[4]その結果として，解析は，実際に検査する必要のない規則を適用したり，直接関係のない無意味な統語構造を作り出したりする

- 言語の特徴を捉えた解析方法を利用する．本研究では，インドネシア語の特徴ある単語をキーワードに用いて文の断片化をまず行なう．次に各断片の意味を解析し，それらを合成させ，意味表現を抽出する方法である．

次章では，われわれのモデルの基本的な枠組について述べる．

4 システムの基本的枠組

自然言語理解の基本的な問題は，文に含まれる様々な曖昧性の解消にあるといえる．インドネシア語も基本的には，他の言語と同じような種類の曖昧性を持っていると考えられるが，インドネシア語に出現すると考えられる曖昧性を以下に示す．

1. 語彙レベル
   (a) 単語の語義による曖昧性
   (b) 格関係による曖昧性
2. 統語レベル
   (a) 係受けによる曖昧性
   (b) 並列性による曖昧性
3. 文脈レベル
   (a) 代名詞の照応による曖昧性
   (b) テンス補完による曖昧性
   (c) 省略補完による曖昧性
   (d) 話者の意図，主題・焦点による曖昧性

高度なインドネシア語理解システムを作成するために，上記の曖昧性解消は同時に行なうべきであると考えるが，本稿では，代名詞の照応，テンスの補完，省略補完などの文脈レベルの解析をまだシステムに導入していないため，文脈レベルの曖昧性解消に関する処理については，簡単に説明することに留める．

以下では，辞書に記述する情報について説明する．

4.1 曖昧性解消のための情報について

意味的曖昧性を解消するために利用する情報には，時制などの文脈に依存した動的な情報と，辞書に記述される統語的制約などの静的な情報がある．

辞書に記述される曖昧性解消のための情報には，語彙共起に関する情報，統語構造に関する情報，意味的制約に関する情報，文脈制約情報，および意味表現への写像に関する情報から構成される．

以下では，これらの情報の記述について説明する．

- 語彙共起に関する情報

"肌が黒い"，"髪の毛が黒い"などのような文は句全体の意味がそれぞれの単語の持つ意味を反映している．このような場合は，各部分の意味を結合させ，全体の意味を計算することができる．これに対して"腹が黒い"のような句が持つ意味は，もはや各部分の意味を反映していない．このような表現は，各部分の意味を計算させるよりも，むしろそれぞれの単語がお互いに共起して一つの語義を持つということを優先的に計算させるべきであろう．本研究では，このような表現に対し，語彙共起(表層的な単語の共起)を制約として語義を与え，意味の計算を優先的に行わせる．例えば，

dia pulang *naik* *kuda* *hijau*
彼 帰る 乗る 馬 緑色
(彼は酔っ払った状態で帰ってきた)

という文は，*naik kuda hijau*の三つの単語が共起して「酔っ払った状態で」という語義を持つものとして記述される．また，*bunga*(花)と*karang*(サンゴ)からの*bunga karang*(スポンジ)や*bunga*と*mulut*(口)からの*bunga mulut*(甘言)のような熟語も同じように扱う[5]．

- 統語構造に関する情報

bunga itu(あの花)と*itu bunga*(あれは花だ)のように単語の順序が異なれば意味も違ってくるということからわかるように意味の抽出には統語規則が一定の役割を果たしていることはいうまでもない．しかし，人間は統語規則に従わない文の意味を理解することも可能であるという事実から考えれば，自然言語理解の計算モデルは，完全に統語規則に束縛される計算モデルではなく，統語規則はあくまでも意味決定を支援する情報の一部として利用されるべきであると考えている．本研究では，以上の考え方を基に優先性を持った統語規則を語義の制約情報として辞書に与える．

統語構造の情報には，動詞パターンを中心としたインドネシア語の基本的文型パターン，名詞句パターン，意味を考慮した前置詞句パターンなどが含まれる．本研究で利用されるインドネシア語の動詞パターンおよび意味を考慮した前置詞句パターンの一部を付録A,Bで示す．

- 意味的制約に関する情報

意味的制約には，isa制約とispartof制約とhasprop制約とableto制約が含まれる．isa制約には，「人間」「道具」「液体」などといった上位概念が制約として使われる．ispartof制約は，例えば，「対象がある概念のある部分である」の上位部分概念の制限に使われる．一方，hasprop制約は，対象の構成要素概念または所有性質の制限に使われる．ableto制約は，対象が行なえるアクション概念の制限に使われる．

- 中間表現への写像に関する情報

この部分では，抽出される文の意味内容を中間表現に変換するための情報が記述される．

- 文脈制約情報について

これには，意図される語義が使われる分野(領域)，状況などの情報が与えられる．

"bunga itu cantik"は，あの花はきれいだ．」または，あの模様柄がきれいだ」の二つの意味を持つ．"bunganya tinggi"も，「花の木が高い」または，「利子が高い」の二つ

[5]英語では，例えば，*not only ... but also*といった表現もこの語彙共起制約で扱う

の意味を持つ．これらの意味の曖昧性は，この単文では区別できない．この曖昧性を解消するには，文脈の情報または発話されている世界の情報が必要である．文脈から抽出されるこれらの情報と，語義の中に記述される情報とを照合することによって語義を決定する．

また，インドネシア語は，英語のように動詞の変形で時制を表現するような言語ではない．"Adi memetik bunga."は，「アディは花を摘む」の現在形であるかまたは，「アディは花を摘んだ」の過去形なのかは文脈を見なければ分からない．このような時制に関する情報は，基本的に文脈から抽出されるとしているが，これが不可能の場合には，辞書が持つ時制のデフォルト情報を利用する．例えば，"terpijak"(踏まれる)の語義には，過去という時制が与えられる．"Bunga itu terpijak"を「あの花は踏まれる」と解釈するのではなく，「あの花は踏まれた」が抽出される．

4.2 優先制約に基づく曖昧性解消法

曖昧性解消法の一つに，語彙項目間の意味的共起関係の整合性によって文の適格性を判定する選択制限による方法がある．選択制限は，曖昧性解消に有効であるとして，多くのシステムはこの考え方を重用している．しかし，これは，選択制限に違反するものをすべて排除することを基本原則としているため，例えば，「食べる」の目的語の選択制限が「食べもの」であるとすれば，「彼は火を食べている」などといった文は，意味的に不適切であるとみなされ，排除される．意味的に少し異常なものでも，文法的であれば，われわれは，解析をやめたり，廃棄したりするのではなく，文脈などを考慮し，一番妥当であると思われる意味を抽出していると考えられる．従って，言語を網羅的に扱うためには，選択制限だけでは不十分である．

選択制限に違反したから排除するというのではなく，文が持つ複数の解釈の確からしさを計算し，それらを比較し，その中で最も確からしい解釈を選択するべきであろう．Y.Wilksによって提案された優先意味論[10]の考え方はそれである．

上記の考え方を参考に，本研究では，選択制限に対して優先性を持たせるという方法を採用する．ここでは，選択制限に対する優先度として，常識的な選択と，ある程度考えられる選択と，非常識的なものを含むそれ以外の選択の三つを用意する．なお，処理は優先性の高い方から行なわれる．

4.3 システムの構成

システムの構成は以下に示す(Fig.1)．

以下では，まず，Fig.1に示されている各部分を説明する．

[形態素解析] 接辞の抽出，畳語の解析を行う

[辞書] インドネシア語単語がそれぞれ持つ語義が登録されている．各語義に記述する情報は，語彙共起情報，統語規則の情報，意味的制約情報と領域制約情報からなる．

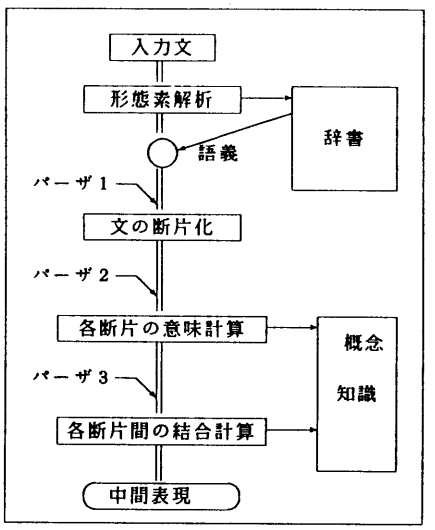

Fig.1 システムの構成

[パーザ1，文の断片化] *-nya*，*yang*，代名詞，動詞，前置詞などの単語をキーワードに用い，文を断片化する．

[パーザ2，各断片の意味計算] 断片化した各単語の語義が持つ情報を基に部分的にそれぞれの意味の妥当性を計算する．

[パーザ3，各断片間の結合計算] パーザ2によって得られた各断片の意味を他の断片との照合を行い，意味を抽出する．

[知識] これには，一般的な概念体系のような静的な知識と，文脈によって得られた，語られている世界，時制などの動的な知識が含まれる．

[中間表現] 抽出された意味表現

解析の基本的な流れは，次のようになる．まず，形態素解析を行う．次に，文を構成する単語の全ての語義を取り出す．語義に記述されている単語のキーワード情報を基に，文を断片化した後，統語規則，意味，文脈制約の情報を用い，各断片の意味を計算する．さらに，動的な文脈情報(時制など)を考慮して断片間の意味を結合し，中間表現を抽出する．

5 解析アルゴリズム

5.1 文の断片化アルゴリズム

文の断片化は，特定の単語をキーワードに用いて文をいくつかの断片に分解することである．インドネシア語の単語の品詞[6]には，名詞，動詞，形容詞，副詞，数詞，不変化詞と代名詞があるとする．これらのうち，動詞，

[6]この単語の品詞の分類は，Balai Pustaka出版のKamus Besar Bahasa Indonesia（インドネシア語大辞典）の分類に従う．

副詞，不変化詞と代名詞（人の名前を含む）に属する品詞の単語が文の断片化のキーワードとして利用される．

不変化詞には，以下のようなものが含まれている．

- yang などの関係詞
- apa,bagaimana などの疑問詞
- telah,lagi などのアスペクトマーカ
- di, ke などの前置詞
- dan, atau などの接続詞
- mungkin, dapat などの様相詞
- pun,lah などの不変化詞

文断片器は，辞書に記述される単語の各語義に与えられている，その単語が文の断片化でキーワードとして利用されるかどうかのキーワード情報を基に，文の単語列を左から順に解析していく．

キーワード情報は，以下の3種類が用意されている．

T1. 自分の前までの単語列を断片にし，自分より後ろの単語からは新しい断片の始まりとする．
例： bunga sakura ↑ *di*↑ ... .

T2. 自分までの単語列を断片にし，自分より後ろの単語からは新しい断片の始まりとする．
例： bunga *Adi*↑ ... .

T3. なにもしないで，(別の単語によって断片にされるまで）そのまま待つ．
例： bunga *sakura* ... .

同じ単語で複数の語義を持ち，それらが異なるキーワード情報を所有する場合は，可能な断片化のすべての組み合わせを抽出するようにする．

例えば：

Komputer bisa menerjemah.[7]

の文を考えてみる．この中の *bisa* という単語には，「できる」と「毒」という語義を持ち，前者はキーワード情報 **T1** に対して後者 **T3** である．結果として以下の2種類の断片の切り方が抽出される．

1. Komputer ↑bisa↑ ↑menerjemah↑.
2. Komputer bisa ↑menerjemah↑.

5.2 各断片の意味計算アルゴリズム

ここでは，文断片器で得られた各断片の意味の計算について説明する．この断片意味解析器は，以下の二つの役割を持つ．

1. 断片が持つ可能な意味の抽出
2. 各解釈への優先度の付与

[7]意味：(コンピュータは翻訳ができる)

各々の単語がすでに各自に語義を持っているので，ここで意味計算が必要となるのは，単語が二つ以上を持つ断片である．断片意味解析器は，まず，二つ以上の単語を持つ断片に対してそれら単語間の修飾のすべての可能な組合せを抽出する．語義に蓄積されている情報と概念知識を利用して意味を計算し，優先度を与える．

具体的な例について説明しよう．以下のような断片が抽出されたとしよう．見やすいように，それぞれに番号を付ける．(なお，説明を簡単にするため，各単語は，一つの語義を持つとする)

harga	bunga	anggrek	Sumatra
値段	花	蘭	スマトラ
1	2	3	4

この四つの単語の断片から以下のように可能な修飾関係をすべて抽出する．

((1 2)(3 4))	...**C1**
((1)((2 3) 4))	...**C2**
((1)((2)(3 4))	...**C3**
((1 (2 3)) 4))	...**C4**
((((1 2) 3) 4))	...**C5**

この場合は，まず，*anggrek* は花の一種であるので，*bunga anggrek*(蘭の花) の **C2** と **C4** に意味優先度を与える．次に，*Sumatra* が，*bunga* かまたは *harga* に修飾するかの処理に移り，*harga Sumatra*（スマトラの値段）か *harga bunga*（花の値段）の意味の妥当性の計算を行う．ここで，「地名」である「スマトラ」と「植物」である「花」を考慮して，**C2** が最終的に選ばれ，意味優先度の高いリストに入れる．意味的優先度が区別できない場合は，曖昧性を持ったままリストに送り込み，断片間の結合計算に委ねる．

5.3 各断片間の結合計算アルゴリズム

断片結合解析器は，以下のような順序で解析が進む．

1. 副詞的な表現を見つけだし，意味的結合を行う．これには，前置詞によって導かれる表現，時間的表現，アスペクト表現，様相表現などが含まれる．
2. 構造的に要求される表現の埋め込みとその意味的妥当性の検査を行う．
3. 文脈情報を参考に省略補完，時制などの処理を行う．
4. 中間表現への写像．

以下では，"Dia memetik bunga di taman." を例に取って断片間の意味結合計算アルゴリズムを簡略化して説明する (文脈を考慮しない)．まず，上の文は，文断片器によって以下のように分解される．

Dia↑	↑memetik↑	bunga	↑di↑	taman↑.
1	2	3	4	5

辞書に記述されている，各単語が持つ語義は，以下のようなものとする[8]．

[8]Sn[i-j]は，i番目の単語からj番目の単語までの第n番目の語義を表す．

Dia :

[dia, 彼, he] ...**S1[1-1]**

[dia, 彼女, she] ...**S2[1-1]**

memetik :

[memetik, 摘む, pick] ...**S1[2-2]**

[memetik, 弾く, play] ...**S2[2-2]**

[memetik, 引く, pull] ...**S3[2-2]**

bunga :

[bunga, 花, flower] ...**S1[3-3]**

[bunga, 利子, interest] ...**S2[3-3]**

[bunga, 模様柄, pattern] ...**S3[3-3]**

taman :

[taman, 庭,garden] ...**S1[5-5]**

解析器は，まず，各断片の意味を計算するが，どの断片も一単語であるため，意味は結果的にそれぞれが持っている語義と同一になる．次に，副詞的な意味を計算するが，この例では，前置詞 *di* の処理に相当する．ここでは，意味を考慮した前置詞パターン[9]を参考に，前置詞 *di* によって構成される全ての意味を抽出する．この場合，*di* と語義 **S1[5-5]** から，パターン **p1a** が適応され，**S1[4-5]** が抽出される．

解析器は，次に，語義に書かれている統語規則を参考に，統語的な結合処理を行なう．以下では，S1[2-2],S1[3-3] の統語構造の情報を例にとって，それらの情報の利用方法について説明する．

[memetik，摘む，pick] の統語的情報：

```
[[  loc:1,                                  ...A1
    o # 1,                                  ...A2
    loc:2,                                  ...A3
    sELF,                                   ...A4
    loc:3,                                  ...A5
    o # 2,                                  ...A6
    loc:4],                                 ...A7
[   p1a([4,1])      → o # 3,                ...B1
    p1b([4,1])      → o # 4,                ...B2
    p2a([1,4])      → o # 5,                ...B3
    p2d([4,1])      → o # 6,                ...B4
    p3a([4,1])      → o # 7,                ...B5
    p4a([4,1])      → o # 8,                ...B6
    p8a([4,1,2])    → o # 9,                ...B7
    ⋱               ⋱                       ⋱
    p9f([4,1])      → o # 13,               ...B11
    tas([2])        → o # 14],              ...B12
[   [p1a,p1b],                              ...C1
    [p2a,p2d],                              ...C2
    [p8a,p8b],                              ...C3
    [p9d,p9e,p9f]]]                         ...C4
```

A4 は，自分自身が現れる場所で，**A2** と **A6** は，それぞれ主語と目的語に相当するもので，これらは，*memetik* の基本動詞パターンを表現したものである．**A1,A3,A5,A7** は，用意された，副詞，ムード，アスペクトなどが随意に現れる場所を示している．

[9]付録Bを参照

B1 ～ **B12** は，副詞的表現，ムード，アスペクト表現などのパターンを表す．例えば，**p1a([4,1])** → **o # 3** の **p1a** は，その表現パターンで，[4,1] の数字は，その表現が出現可能な場所を表す．ここでは，**A7** の **loc:4** と **A1** の **loc:1** が，**p1a** が出現可能な場所となる．なお，数字の順序は，一般にもっともよく使われる方から順に並べたものである．また，**o # 3** は，**p1a** の意味内容を示す．

C1 ～ **C4** は，**B1** ～ **B12** の共起制限である．例えば，**C1** は，パターン **p1a** と **p1b** が，同時に出現することがないことを表している．

次に，S1[3-3]**bunga** の統語構造の情報には，

```
[[   loc:1, sELF, loc:2 ] ,
 [   p1a([2]) → o # 1,
     p9e([2]) → o # 2 ] ,
 [   ]]
```

が記述されているとする．もし，bf S2[3-3] と S3[3-3] が S1[3-3] と同じ統語構造を持つのであれば，*bunga di taman* の構造的係り受けは，次のように展開される．

Dia	memetik	bunga	di taman.
S1[1-1]	S1[2-2]	S1[3-3]	S1[4-5]
S2[1-1]	S2[2-2]	S2[3-3]	
	S3[2-2]	S3[3-3]	
		S1[3-5]	{S1[3-3]+S1[4-5]}
		S2[3-5]	{S2[3-3]+S1[4-5]}
		S3[3-5]	{S3[3-3]+S1[4-5]}

また，**S1[2-2],S2[2-2],S3[2-2]** に対しても同じように展開される．

次に，パーザ2によって各断片の意味が計算されるが，その前に，統語構造の情報を参考に各断片の配置を割り当てる．結果として以下のような2種類の組合せが得られる．

1. o # 1 = $S_{1\ 2}$[1-1], sELF = $S_{1\ 3}$[2-2], o # 2 = $S_{1\ 3}$[3-3], o # 3 = S1[4-5][10].

2. o # 1 = $S_{1\ 2}$[1-1], sELF = $S_{1\ 3}$[2-2], o # 2 = $S_{1\ 3}$[3-5]

S1[4-5] は，$S_{1\ 3}$[2-2] または $S_{1\ 3}$[3-3] に修飾するかという曖昧性である．これは，一般に文脈の情報から決定されるが，文脈の情報がなければ，デフォルトとして優先的に動詞の方に修飾するように処理する．従って，この例の場合は，後者が優先的に選択される．次に，意味制約情報を利用して計算を行なう．S2[2-2] と S3[2-2] は，他の要素の意味との整合性がないため，それに対応する優先度が抽出される．S1[2-2] は，o # 1 が「人間」で，o # 2 が「花」か「枝」の部類概念が指定されるとすれば，最終的に，優先的に S1[2-2] が選択される．また，S1[1-1] と S2[1-1] については，文脈からしか曖昧性を解消できないので，曖昧なまま処理される．

S1[2-2] の統語構造に対応して中間表現への写像情報が与えられているので，最終的に，(([memetik,摘む,pick] , agent, [dia, 彼,he]) ([memetik, 摘む,pick] , object, [bunga, 花,flower]) ([memetik, 摘む,pick] , location, [taman, 庭,garden])) のような意味が抽出される．

[10]S1[4-5]は，場所格 **p1a** に属している

6 おわりに

本研究では，インドネシア語の特徴ある単語をキーワードにして文を断片化し，語義辞書に書かれている統語規則の情報，意味情報などを手がかりに，分割された各断片の曖昧性を段階的に解消し，文の意味内容を直接抽出する解析方法を提案した．語彙共起，統語規則，意味と文脈の情報を一括して各語義に記述することによって言語表現の形態論と統語論と意味論との関係を統一的かつ明示的に表現することができた．文を構成する単語の語義の数だけで解析処理を行なうことにより，膨大な数の規則を扱う必要がある，従来の解析方法に見られるような直接関係のない無意味な構造を作り出したりする問題は解決される．

文を断片化するやり方で文解析を行なう例としてWilksの研究[10]があるが，複雑な文の解析は困難であることが認識されている．これは，文の統語的な関わりをすべて意味解析アルゴリズムの中に暗黙に取り込んでいることが解析の限界の原因になっていると思われる．これに対して我々は，構文規則を語義の制約情報として与え利用することによって上記の問題を解決できると考える．

しかし，各語義に個別にそれぞれの統語規則，意味情報などの詳細な情報を与えることによって語義を独立的に開発することができ，管理も比較的容易にできる一方，語義辞書がかなり大きくなるという問題点がある．これは今後の課題として残すが，解決策として，語義が持っている情報を体系化して，共通に所有する情報を継承的に記述するなどといった方法などが考えられる[3]．また，より効率よく文の断片化を処理するには，インドネシア語の特徴をより徹底的に研究し洞察する必要がある．本研究の解析方法は，基本的にインドネシア語を処理するために考えたものではあるが，比較的語順の自由な文法を持つ言語にも利用できると思われる．

参考文献

[1] 辻井潤一．機械翻訳のための文法とその問題点．，第一回「大学と科学」シンポジウム，「日本語特性と機械翻訳」予稿集，1987．

[2] 田中穂積．自然言語解析の基礎．産業図書，1989．

[3] D. Flickinger, C. Pollard, and T. Wason. Structure-sharing in lexical representation. In *The Procs. of the 23th Annual Meeting of the ACL, Chicago*, pages 262–267, 1985.

[4] H. Maruyama. Structural disambiguation with constrain propagation. In *The Procs. of the 28th Annual Meeting of the ACL*, pages 31–38, 1990.

[5] C.S Mellish. Incremental semantic interpretation in a modular parsing system. In K. Sparck-Jones and Y. Wilks, editors, *Automatic Natural Language Parsing*, Ellish Horwood, 1983.

[6] Manabu Okumura and Hozumi Tanaka. Towards incremental disambiguation with a generalized discrimination network. In *The Procs. of AAAI-90*, pages 990–995, 1990.

[7] Chuck Rieger and Steve Small. Word expert parsing. In *The Procs. IJCAI-29, Tokyo*, pages 723–728, 1979.

[8] Ivan Sag and Carl Pollard. *Head-Driven Phrase Structure Grammar: An Informal Synopsis.* Technical Report, Center for the Study of Language and Information,Stanford, 1987. Technical Report CSLI-87-79.

[9] Vises Vorasucha. *A Study on Thai Language Analysis Based on Head Grammar.* PhD thesis, Tokyo Institute of technology, 1989.

[10] Y. Wilks. An artificial intelligence approach to machine translation. In R.C. Schank and K.M. Colby, editors, *Computer Models of Thought and Language*, pages 114–151, W.H. Freeman adn Company, 1973.

Appendix

A インドネシア語の動詞パターン

名詞句＋動詞＋名詞句
名詞句＋動詞＋名詞句＋動詞
名詞句＋動詞＋名詞句＋動詞＋名詞句
名詞句＋動詞＋名詞句＋形容詞句
名詞句＋動詞＋名詞句＋名詞句
名詞句＋動詞＋名詞句＋ bahwa節
名詞句＋動詞＋疑問節
名詞句＋動詞＋ bahwa節
名詞句＋動詞
名詞句＋動詞＋形容詞句
名詞句＋動詞＋動詞
名詞句＋動詞＋動詞＋名詞句

B 意味を考慮した前置詞句パターン

ただし，パターンの記号として，例えば，1aの場所・方向を示す「ｄｉ＋名詞句」は**p1a**で表すとする．

1. 場所・方向の指示
   (a) di＋［場所表現］
   (b) dari＋［場所表現］
   (c) ･･･
2. 時間の指示
   (a) di＋［時間表現］
   (b) ･･･
3. 使用道具の指示
   (a) dengan＋［道具表現］
   (b) tanpa＋［道具表現］
   (c) ･･･
4. ⋱