修 士 論 文

インドネシア語の形態素解析に関する研究

指導教官　佐藤理史 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

Muizzul Hidayat

1997 年 5 月 15 日

目 次

第 1 章

序論

形態素解析は自然言語処理の基本技術の一つであるにもかかわらず，インドネシア語の形態素解析システムはまだ十分に研究されていない．インドネシア語の形態素解析のむずかしさは，単語に接辞を付加すること，および，接辞付加後の品詞変化にある [1]．

インドネシア語の単語は，大きく，基語と構成語に分けられる．このうち構成語は，基語を反復したり（重複語），基語や重複語に様々な接頭辞や接尾辞を付加することによって作られる単語である．構成語に対して，さらに接頭辞や接尾辞を付加することによって新しい構成語が作られることもある．このようにインドネシア語の語構成に関する規則は，かなり複雑である．このため，インドネシア語の形態素解析の実現においては，この構成語をどう取り扱うかが問題となる．

英語や日本語を対象とした形態素解析は，比較的よく研究されている．英語は，単語間に空白が存在するとともに，語形変化がある単語が少なく，基本的に，動詞，形容詞，副詞，名詞にしか接尾辞が付かないため，単語の原形のみを辞書に登録しておき，接尾辞に関する規則を併用することによって，比較的簡単に形態素解析を実現できる．これに対して，日本語は，単語間に空白が存在しないため，形態素解析の主な処理は，文を単語に分割することであり，多くの場合，すべての形態素をあらかじめ辞書に登録しておく方法が用いられる．

本研究で対象とするインドネシア語は，英語や日本語とは違う特徴を持った言語であり，語構成に関するかなり複雑な規則を持っている．このため，形態素解析としてどのような方法が適切であるか，不明である．

インドネシア語の形態素解析を実現する方法は，大きく２つの方法に大別できる．第一

の方法は，全ての単語を辞書に登録しておく方法 [2] である．この方法をとる場合，形態素解析のアルゴリズムは単なる辞書引きでよく，非常に簡単になるが，膨大な数の単語をどうやって網羅的に収集するかということが問題となる．

第二の方法は，基本となる単語（基語）だけを辞書に登録しておき，語構成に関する規則を併用する方法 [1] である．この方法では，形態素解析のアルゴリズムは多少複雑になるが，辞書のサイズは小さく押えることができる．しかしながら，この方法では，語構成に関する規則をどうやって獲得するかということが問題となる．

本研究では，第二の方法で問題となる，語構成に関する規則の獲得に，第一の方法のために作成された既存の辞書を利用することを検討する．ここで用いる辞書は，CICC プロジェクトで作成されたＩＭＤ（Indonesian Master Dictionary）である．この辞書は，すべての単語を登録するという立場に立って作られているため，基語の他に，構成語を含んでいる．まず，辞書中に含まれる構成語を分解し，接辞のリストと語構成規則を獲得する。次に，こうして得られた知識を形態素解析時に利用し，ＩＭＤに含まれていない単語も解析できるような，より広いカバレジを持った形態素解析システムを実現する．

本論文の構成は，以下の通りである．まず，第２章で，インドネシア語の形態素の特徴と，インドネシア語の形態素解析の現状について述べる．第３章では，ＩＭＤから接辞リストや語構成規則をいかにして獲得するかについて述べる．第４章では，本研究で作成した形態素解析システムについて述べ，第５章では，実験と評価について述べる．最後に第６章で，結論を述べる．

第 2 章

インドネシア語の形態素解析の現状

本章では，まず本研究の内容を理解するために必要なインドネシア語の形態素の特徴について説明を行なったのち，本研究の対象であるインドネシア語の形態素解析の現状について述べる．さらに従来の方法の問題点も示す．

2.1 インドネシア語の形態素の特徴

インドネシア語の単語は，大きく，基語と構成語に分けられる．基語は，接辞が付加されておらずかつ重複もしない，単語として最も小さな単位である．基語はあらゆる品詞にわたってみられる．構成語は，基語を重複したり，基語や重複語に様々な接辞を添えることによってできる単語である．その例を図 2.1に示す．インドネシア語の単語の構成を図 2.2，図 2.3に示す．

基語に接辞を付加することによって，以下の２つが実現される．

1. １つの基語から様々な意味の語がつくり出せる．

2. １つの基語から様々な品詞の語がつくり出せる．

一定の接辞は，基語から一定の傾向の意味や品詞を持つ構成語をつくり出す [3]．したがって，構成語の場合には，その単語の語構成からその単語の品詞や用法が推定できると考えられる．

基語：	**lihat**	**動詞**	**見る**
構成語	ter*lihat*	動詞	ふと見る
	ke*lihat*an	動詞	見える
	pe*lihat*	名詞	予見者
	peng*lihat*an	名詞	視力
	me*lihat*	動詞	ながめる
	me*lihat*i	動詞	よく見る
	me*lihat*kan	動詞	見る
	di*lihat*	動詞	見られる
	per*lihat*kan	動詞	見せる
	me*lihat*–lihat	動詞	見物する
	ber*lihat*–lihatan	動詞	お互いに見会わす
	ke*lihat*annya	形容詞	…のようである
	⋮	⋮	⋮

図 2.1: 単語の例

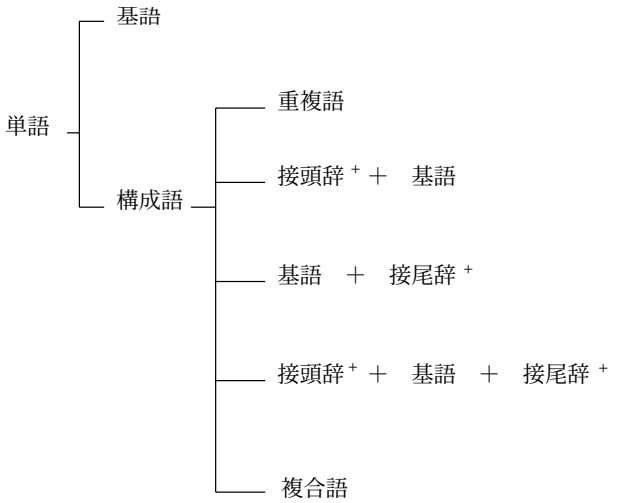


図 2.2: 単語の構成１

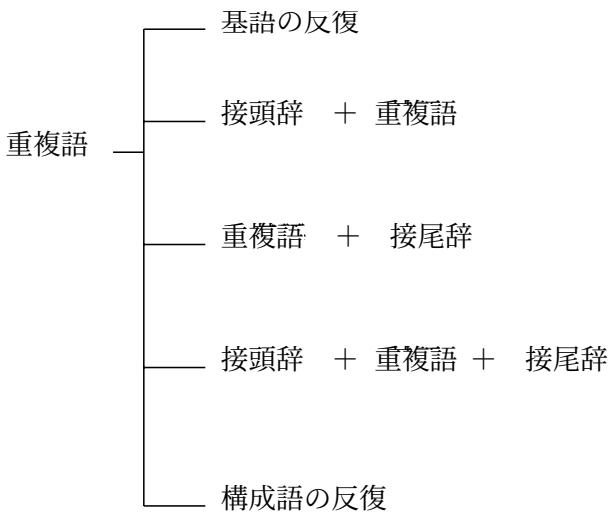


図 2.3: 単語の構成２

2.2 インドネシア語の形態素解析の現状

インドネシア語の形態素解析の方法として，現在までに，次の２つの方法が提案されている．

2.2.1 Yusuf の方法

この形態素解析手法は，屈折形のある単語から基語を抽出するために音韻論的な処理および形態素処理を併合したものである [1]．この音韻的な処理では，屈折接辞と基語を音韻単位に分解し，屈折接辞の音韻変化規則を用いて，活用を単なる屈折接辞と基語の接続として捉える．

例えば，入力文が ([meng] ＋ [kurang ＋ action]) とした場合，この手法では，

([Affix] ＋ [Root Word ＋ Semantic]) −→ [Word ＋ New Semantic]

という形式の形態素規則と音韻変化規則が用いられ，その解析結果は以下のようになる．

([meng] ＋ [kurang ＋ action]) −→ [me**ng**urang ＋ active]

しかし，同じような場合で屈折しない場合も存在する．例えば，以下の例では，k が ng に屈折しないのが正しい．

表 2.1: ＩＭD辞書の単語数

	数
基語	16,506
構成語	14,982
計	31,488

([meng] ＋ [kaji]) −→ [meng<u>k</u>aji]

2.2.2 全ての語を辞書に登録する方法

この方法は，全ての単語（基語，重複語，構成語）を辞書に登録しておくという方法であり，形態素解析は，単なる辞書引きとなる．この立場に立って作成された代表的な辞書に，ＣＩＣＣプロジェクトのＩＭDがある．表 2.1に，ＩＭDの基語数および構成語数を示す．その語彙は 3 万語強であり，十分とは言えない．

第 3 章

既存辞書からの知識獲得

本章では，まず，既存の辞書からの知識獲得の考え方を述べた後，既存辞書として利用したＩＭＤの構成と，その辞書からの接辞の獲得，および，語構成規則の獲得について述べる．

3.1　既存辞書からの知識獲得の考え方

インドネシア語では，様々な接辞を付加できる単語がかなり多く存在するため，全ての単語（構成語）を網羅的に辞書登録することは困難である．そのため，限られた単語数で，カバレジをあげる方法が必要になる．

そのために必要となるものは，(1) 接辞のリスト，(2) どのような接辞がどのような語に接続するかという接続情報，(3) ある接辞が付加された場合，その語は全体としてどういう品詞となるかという語構成規則，の３つである．このような知識は，どこから獲得することができるであろうか．

これらの知識を，すべて人手で作るという方法も考えられるが，本研究では，既存の辞書からこれらの知識を半自動的に獲得することを考える．すなわち，辞書中に存在する構成語を形態素に分解して，接頭辞，接尾辞，語構成規則などを獲得する．得られた知識を形態素解析時に利用し，よりカバレジの広い形態素解析システムを実現するというのが本研究の基本的なアプローチである．

その理由を以下に述べる．

1. インドネシア語は，基語に様々な接頭辞，接尾辞を添えることによって，基語から

様々な構成語をつくり出せるという特徴を持っている．特定の接辞の付加によって生成される構成語は，特定の品詞をとるという傾向をもっている．したがって，構成語の品詞は，その語構成から推定できると考えられる．

2. 現在，整備された辞書は構成語をそのままに登録している．しかし，その辞書は単語数が少なく，不十分である．

3.2 ＩＭＤの構成

本研究では，知識獲得の対象として使用する既存の辞書として，１９９５年３月に開発された Indonesian Master Dictionary （ＩＭＤ）を用いた．

ＩＭＤでは，単語は基語ごとに分類されており，以下のようなフォーマットで記述されている．

基語部			: 基語
			: 語の番号（６桁）
			: バリエーション
			: 基語の概念
品詞部			: 品詞名
	対応の語		: 単語
		能動態情報	: 能動態コード
		受動態情報	: 受動態コード
		その他情報	: その他コード
	概念情報		: 概念記述
		意味情報	: 意味素性
		概念番号	: 概念番号

基語 *adaptasi* に対するＩＭＤの記述を図3.1に示す．この基語に対して，*adaptasi*，*pengadaptasian*，*beradaptasi*，*mengadaptasikan* という４つの単語が定義されている．このうち，最初の一つは，基語それ自身であり，残りの３つは，その基語から作られる構成語である．

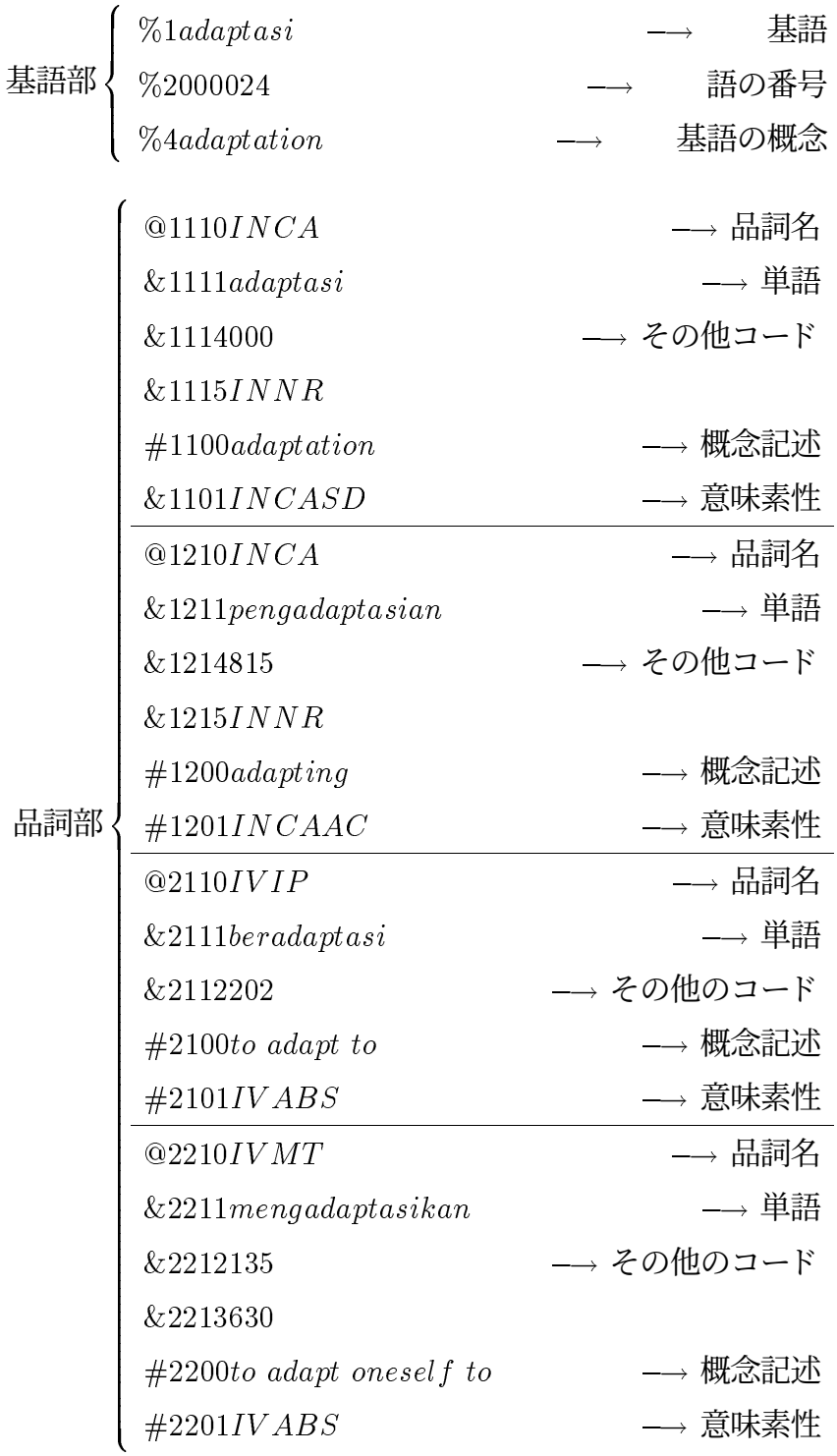

基語部
%1adaptasi ⟶ 基語
%2000024 ⟶ 語の番号
%4adaptation ⟶ 基語の概念
品詞部
@1110INCA ⟶ 品詞名
&1111adaptasi ⟶ 単語
&1114000 ⟶ その他コード
&1115INNR
#1100adaptation ⟶ 概念記述
&1101INCASD ⟶ 意味素性
@1210INCA ⟶ 品詞名
&1211pengadaptasian ⟶ 単語
&1214815 ⟶ その他コード
&1215INNR
#1200adapting ⟶ 概念記述
#1201INCAAC ⟶ 意味素性
@2110IVIP ⟶ 品詞名
&2111beradaptasi ⟶ 単語
&2112202 ⟶ その他のコード
#2100to adapt to ⟶ 概念記述
#2101IVABS ⟶ 意味素性
@2210IVMT ⟶ 品詞名
&2211mengadaptasikan ⟶ 単語
&2212135 ⟶ その他のコード
&2213630
#2200to adapt oneself to ⟶ 概念記述
#2201IVABS ⟶ 意味素性

図 3.1: ＩＭＤの記述例

表 3.1: ＩＭＤの語彙

	数	品詞付き	品詞なし
基語	16,506	14,097	2,409
構成語	14,982	14,982	0
計	31,488	29,097	2,409

3.3 ＩＭＤの変換

本研究では，まず，ＩＭＤから必要な基語，構成語，品詞を抽出して，以下のようなフォーマットに変換した．

単語　単語の品詞$^+$　基語　基語の品詞$^+$

ここで+は，１つ以上の繰り返しを意味する．なお，ＩＭＤにおいては，全ての基語が品詞部で定義されているとは限らない．基語部には品詞に関する情報はないため，このような場合には，基語の品詞が得られない．その場合は，UNKNOWN として登録することとした．

例えば，図 3.1の記述は，以下のように変換される．

```
adaptasi          INCA   adaptasi   INCA
pengadaptasian    INCA   adaptasi   INCA
beradaptasi       IVIP   adaptasi   INCA
mengadaptasikan   IVMT   adaptasi   INCA
```

本研究では，こうして得られた変換後の辞書から，各種の知識獲得を行なった．

変換後の辞書の語彙を表 3.1に示す．また，それぞれの単語に品詞がいくつ付けられているかを調べた結果を表 3.2に示す．また，品詞の分布を表 3.3に示す．

3.4 接辞の獲得

インドネシア語では，接辞は，ある基語あるいは構成語に付加される形態素で，付加することによって生成される構成語の意味や品詞に影響を直接に与える．ここでは，接辞を

表 3.2: ＩＭＤの品詞の個数

	品詞の個数			計
	1	2	3	
基語	13,829	264	4	14,097
構成語	14,928	54	0	14,982
計	28,757	318	4	29,079

表 3.3: ＩＭＤの品詞の分布

品詞	基語	構成語	計
名詞	10,730	4,992	15,722
代名詞	28	5	33
形容詞	2,965	425	3,390
動詞	343	9,272	9,615
副詞	101	201	302
数詞	28	39	67
冠詞	3	0	3
間投詞	30	1	31
社交カテゴリー	5	0	5
接続詞	58	32	90
前置詞	34	15	49
助動詞	29	0	29
限定詞	12	54	66
不変化詞	3	0	3
計	15,036	14,369	29,405

ＩＭＤ（変換後）から獲得する方法とその結果について述べる．

3.4.1 接頭辞の獲得

接頭辞は，ある基語あるいは構成語の前に接続し，新たに構成語をつくり出す機能をもっている一つの形態素である．基語あるいは構成語には複数の接頭辞が付加される場合もある．

接頭辞の獲得アルゴリズムは，大きく２つのステップに分けられる．

第一のステップ（ステップ１）は，ＩＭＤに含まれるそれぞれの構成語から，接頭辞の列を切り出すことを行なう．先に述べたように，変換語のＩＭＤの各単語には，その単語の基となる基語が付けられている．このため，以下のアルゴリズムで，接頭辞の列を切り出すことができる[1]．

構成語から接頭辞列を切り出すアルゴリズム

1. ２つの文字列（構成語と基語）が与えられる．
2. もし構成語中に基語が含まれているならば，次の操作を行なう．
   (a) 構成語中のマッチしていない前部分文字列が存在するならば，その前部分文字列を，接頭辞列とする．
3. ２の条件を満たさない場合で，構成語中に，先頭文字を取り除いた基語が含まれているならば，次の操作を行なう．
   (a) 構成語中のマッチしていない前部分文字列が存在するならば，その前部分文字列を，接頭辞列とする．

インドネシア語では，接頭辞と基語が接続した場合，基語の先頭の１文字が脱落する場合がある．上記アルゴリズムの 3. は，これに対応した処理である．

上記のアルゴリズムを，ＩＭＤに含まれる構成語に適用することによって，接頭辞列の集合を得ることができる．

[1] インドネシア語では，一つの基語に複数の接頭辞が付加されることがあるため，ここで得られるものは，接頭辞ではなく，一般に接頭辞の列となる．

第二のステップ（ステップ２）では，こうして得られた接頭辞列の集合から，接頭辞の集合を求めることを行なう．これを機械的に実現するために，以下の仮説を採用する．

接頭辞は，単独で付加される例が存在する．

いま，ステップ１で得られた集合の中に文字列 $abcd$ が含まれる場合を考えよう．これが，ab と cd という２つの接頭辞の列であるためには，ab と cd がそれぞれ，ステップ１で得られた集合に含まれていなければならない，ということを上記の仮説は要請する．もし，その条件を満たさない場合は，$abcd$ を一つの接頭辞と考える．

接頭辞列のリストから接頭辞のリストを得る具体的なアルゴリズムは，以下で与えられる．なお，ここでは，接頭辞は２文字以上からなる文字列であるという条件を加えた．

接頭辞列のリストから接頭辞のリストを求めるアルゴリズム

1. 接頭辞列のリストを短い順にソートする．
2. 接頭辞のリストを空とする．
   Prefix := ();
3. 接頭辞列のリストの各要素 x に対して以下に繰り返す．
   (a) もし **x** の長さが（１＜ | x | ≤ ３）文字であるならば，**x** を **Prefix** に含める．
   (b) もし **x** が現在の **Prefix** に含まれる要素に分割できるならば，何もしない．
   (c) そうでなければ，**x** を **Prefix** に含める．
4. **Prefix** が求める接頭辞のリストである．

本方法を，実際にＩＭＤに適用したところ，ステップ１において，61 個の接頭辞列が得られ，ステップ２において，28 個の接頭辞が得られた．このうち，ステップ２における分割結果を表 3.4に示す．61 個の接頭辞列のうち，54 個 (89%) は正しく分割された．最終的にシステムが提案した 28 個の接頭辞を表 3.5に示す．このうち２１個が正しい接頭辞であり，７個は誤りであった．誤った原因は以下に２つのケースに分けられる．

表 3.4: 接頭辞列の分割結果

接頭辞列中の接頭辞数	正解数	誤り数	計
1	21	0	21
2	31	6	37
3	2	1	3
計	54	7	61

1. 接頭辞列中に単独の接頭辞が存在しなかった.
   例：接頭辞列 bersi, pemer, perike はそれぞれ単独の接頭辞 si, mer, peri が存在しなかった.

2. 屈折した接頭辞である.
   例：接頭辞列 menge, penge, menye, penye はそれぞれ ng, ny が k, s に変化する.

こうして得られた結果に基づき，インドネシア語の接頭辞リストを作成した．このリストを表3.6に示す．このリストに含まれる 33 個の接頭辞のうち，自動的に獲得できたものは 21 個であり，残りの 12 個（*が付けられたもの）は人手で追加した．なお，それぞれの接頭辞の標準形は，すべて人手で付加した.

3.4.2 接尾辞の獲得

接尾辞は，ある基語あるいは構成語の後に接続し，新たな構成語を作り出す形態素である．接頭辞と同じように，インドネシア語では，基語あるいは構成語に複数種類の接尾辞が付加される場合がある.

ＩＭＤから接尾辞を獲得する方法は，接頭辞の獲得方法とほぼ同じである．但し，接頭辞は２文字以上であることを条件としたが，接尾辞は１文字からなる場合も許す.

接頭辞の獲得の場合と同様に，接尾辞の獲得を実際にＩＭＤに適用したところ，ステップ１において，49 個の接尾辞列が得られ，ステップ２において，39 個の接尾辞が得られた．このうち，ステップ２における分割結果を表3.7に示す．49 個の接頭辞列のうち，46 個 (94%) は正しく分割された．最終的にシステムが提案した 39 個の接頭辞を表3.8および表3.9に示す．これらは，全て正しい接尾辞であった.

表 3.5: システムが提案した接頭辞の評価

システムが提案した接頭辞	頻度	評価	正解	原因
be	93	○	be	–
di	1	○	di	–
ke	1,073	○	ke	–
le	2	○	le	–
me	1,276	○	me	–
pe	402	○	pe	–
se	266	○	se	–
te	18	○	te	–
bel	1	○	bel	–
ber	2,079	○	ber	–
mem	1,210	○	mem	–
men	1,462	○	men	–
pel	6	○	pel	–
pem	407	○	pem	–
pen	544	○	pen	–
per	552	○	per	–
ter	853	○	ter	–
meng	1,858	○	meng	–
meny	594	○	meny	–
peng	554	○	peng	–
peny	231	○	peny	–
bersi	1	×	ber + si	si が単独で存在しない
menge	41	×	me + ke	ng が k に変化
menye	21	×	me + se	ny が s に変化
pemer	2	×	pe + mer	mer が単独で存在しない
penge	19	×	pe + ke	ng が k に変化
penye	3	×	pe + se	ny が s に変化
perike	1	×	peri + ke	peri が単独で存在しない

表 3.6: 最終的に作成した接頭辞リスト

接頭辞	接頭辞の標準形
be	ber
bel	ber
ber	ber
de^*	de
di	di
je^*	je
kau^*	kau
ke	ke
ku^*	ku
le	le
me	me
mem	me
men	me
meng	me
$menge^*$	$me + ke$
meny	me
$menye^*$	$me + se$
mer^*	mer
pe	pe
pel	pe
pem	pe
pen	pe
peng	pe
$penge^*$	$pe + ke$
peny	pe
$penye^*$	$pe + ke$
per	per
$peri^*$	$peri$
se	se
si^*	si
te	ter
tel^*	ter
ter	ter

表 3.7: 接尾辞の分割結果

接尾辞列中の接尾辞数	正解数	誤り数	計
1	39	0	39
2	4	3	7
3	3	0	3
計	46	3	49

こうして得られた結果に基づき，インドネシア語の接尾辞リストを作成した．このリストを表3.10および表3.11に示す．ここで人手の追加した接尾辞（*）は，７個である．その内３個が分割の誤りを修正して，新たに接尾辞として追加したものである．

3.5 辞書中の構成語の分解

こうして得られた接頭辞リストと接尾辞リストを用いることによって，ＩＭＤに含まれる構成語を形態素に分解することが可能となる．具体的には，以下のアルゴリズムによって，構成語を分解する．

構成語の分解アルゴリズム

1. ２つの文字列（構成語と基語）が与えられる．
2. もし構成語中に基語が含まれているか，あるいは，先頭文字を取り除いた基語が含まれているならば，次の操作を行なう．
   (a) 構成語中のマッチしていない前部分文字列を切り出し，接頭辞リストを用いて，接頭辞の列に分解する．
   (b) 構成語中のマッチしていない後部分文字列を切り出し，接頭辞リストを用いて，接頭辞の列に分解する．
   (c) 分解結果を出力する．

上記のアルゴリズムを用いて，それぞれの構成語に対して，以下のような情報を得る．

表 3.8: システムが提案した接尾辞の評価

システムが提案した接尾辞	頻度	評価	正解	原因
f	4	○	f	–
i	773	○	i	–
s	6	○	s	–
ah	1	○	ah	–
al	3	○	al	–
an	3,985	○	an	–
at	2	○	at	–
er	2	○	er	–
if	3	○	if	–
ik	5	○	ik	–
is	14	○	is	–
me	9	○	me	–
ni	1	○	ni	–
si	5	○	si	–
wi	1	○	wi	–
asi	2	○	asi	–
iah	1	○	iah	–
ika	1	○	ika	–
ita	2	○	ita	–
kan	2,261	○	kan	–
lah	1	○	lah	–
man	1	○	man	–
nya	62	○	nya	–
sis	1	○	sis	–
tal	4	○	tal	–
tik	2	○	tik	–
wan	8	○	wan	–
etik	1	○	etik	–
gram	1	○	gram	–
itis	1	○	itis	–

表 3.9: 接尾辞の獲得結果

システムが提案した接尾辞	頻度	評価	正解	原因
logi	1	○	logi	–
ogen	1	○	ogen	–
olog	1	○	olog	–
onal	2	○	onal	–
oner	1	○	oner	–
sida	1	○	sida	–
wati	3	○	wati	–
grafi	1	○	grafi	–
istis	1	○	istis	–

1. 構成語
2. 構成語の品詞
3. 接頭辞表層形：接頭辞標準形（の並び）
4. 基語表層形：基語標準形
5. 基語の品詞
6. 接尾辞表層形：接尾辞標準形（の並び）

例を表 3.12に示す.

以下で述べる，語構成規則の獲得と接続情報の獲得は，こうして得られた構成語の分解結果を用いる.

3.6 語構成規則の獲得

ここでいう語構成規則とは，ある単語にある接辞が付加されることによって構成されている構成語の品詞を，元の単語と付加された接辞から決定する規則のことである．前節で

表 3.10: 接尾辞標準形

接尾辞	接尾辞の標準形
ah	ah
al	al
an	an
asi	asi
at	at
$atif^*$	$atif$
er	er
etik	etik
f	f
gram	gram
grafi	grafi
i	i
iah	iah
if	if
ik	ik
ika	ika
is	is
$isasi^*$	$isasi$
ita	ita
itis	itis
istis	istis
kah^*	kah
kan	kan
ku^*	ku
lah	lah
logi	logi
man	man

表 3.11: 接尾辞標準形

接尾辞	接尾辞の標準形
me	me
*mu**	*mu*
nya	nya
ni	ni
ogen	ogen
olog	olog
*ologi**	*ologi*
onal	onal
oner	oner
*pun**	*pun*
s	s
si	si
sida	sida
sis	sis
tal	tal
tik	tik
wan	wan
wati	wati
wi	wi

表 3.12: 構成語の分解例

構成語	pengadaptasian
構成語の品詞	INCA
接頭辞表層形：接頭辞標準形（の並び）	peng:pe
基語表層形：基語標準形	adaptasi:adaptasi
基語の品詞	INCA
接尾辞表層形：接尾辞標準形（の並び）	an:an

述べた構成語の分解結果は，その実例となっているわけであるから，これを一般化することによって，語構成規則を獲得することができる．

いくつかの予備的調査を行なった結果，以下の方針をとることとした．

- 語構成規則の形式としては，以下の３つのみを採用する．

$$\begin{array}{lllll} \text{接頭辞} + & X & & \rightarrow & \text{構成語品詞} \\ & X & + \text{接尾辞} & \rightarrow & \text{構成語品詞} \\ \text{接頭辞} + & X & + \text{接尾辞} & \rightarrow & \text{構成語品詞} \end{array}$$

  ここで，接頭辞，接尾辞には，それぞれ具体的な接辞が入る．また，X のところには，任意の基語又は構成語が入り得るものとする．

- 品詞は，大分類を採用する[2]．

語構成規則に含まれる接頭辞，接尾辞の数を，それぞれ１つまでに制限したのは，インドネシア語の構成語の品詞は，最後に付けられた接辞（つまり，一番先頭の接頭辞か一番末尾の接尾辞）によってほぼ決定されるからである．また，品詞の細分類を使用せず，大分類を使用したのは，細分類を使用すると語構成規則の数が非常に多くなるためである．

このような方針に基づいて一般化を行なうと，表 3.12 からは，以下のような語構成規則が得られる．

$$\text{pe} + X + \text{an} \rightarrow \text{NOUN}$$

このように，構成語の分解結果を一般化することによって，語構成規則を獲得することができるが，これを多数の構成語の分解結果に対して行なうと，条件部（矢印の左側）が同一で，帰結部（矢印の右側）だけが異なる規則が生成されることがある．これは，構成語の品詞が接辞だけからは一意に定まらないことを意味する．ここでは，条件部が同一の規則が複数ある場合は，それらの実例の数を数え，その頻度が 全体 80 % 以上の語構成規則が存在した場合は，それが基本規則と考え，他のものを削除することを行なった．

最後に，条件部が同一の規則が複数存在する場合は，以下のように，帰結部をマージすることを行ない，これを最終的な語構成規則とする．

[2]付録 A の品詞分類表を参照のこと．

表 3.13: 語構成規則の概要

品詞の数	規則の数
1	62
2	6
3	5
4	2
5	0
6	0
7	0
8	1
計	76

条件部 → 品詞 1
条件部 → 品詞 2
⇓
条件部 → 品詞 1, 品詞 2

以上説明した方法によって，ＩＭＤから 76 個の語構成規則を獲得した．それらの規則の帰結部にいくつ品詞が含まれているかを表 3.13 に示す．この表 3.13から分かるように，複数の品詞を帰結する語構成規則（つまり，品詞の曖昧性を発生させる規則）が１４個（18 %）ある．

3.7 接続情報の獲得

インドネシア語では，ほとんどの接辞はほとんどの基語に接続可能であるため，ある特別な場合を除いて，接続はすべて可能であるとしてよい．

この特別な場合とは，接頭辞の接続によって，基語の先頭文字が脱落する場合である．例えば，接頭辞 *me* が *p* で始まる基語，例えば *pakai* に接続する場合，接頭辞は *mem* となり *p* が脱落して，*memakai* となる．

そこで，それぞれの接頭辞（表層形）に対して，それがどんな先頭文字を持つ基語に接

続するか，あるいは，どんな先頭文字を持つ基語に接続する場合に先頭文字が脱落するかを，構成語の分解結果から抽出する．例えば，表 3.12 からは，接頭辞表層形 *peng* が基語 *adaptasi* に接続し，その先頭文字は脱落しないので，

peng　+a

を得る．ここで+a は，先頭文字が *a* である基語に接続し，かつ，それが脱落しないことを意味する．先の *memakai* からは，

mem　−p

を得る．ここで−p は，先頭文字が *p* である基語に接続し，かつ，それが脱落することを意味する．

第 4 章

形態素解析システムの作成

本章では，本研究で作成したインドネシア語の形態素解析システムについて述べる．まず，システムが使用する辞書およびテーブルについて述べ，次に，形態素解析アルゴリズムについて述べる．

4.1 辞書およびテーブル

本形態素解析システムの構成を図4.1に示す．本システムでは，辞書，接続テーブル，語構成規則テーブルの３つの知識源を用いる．

4.1.1 辞書

本システムでは，それぞれの形態素に関する個別的な知識は，全て辞書の中に記述する．辞書の記述形式は次のとおりである．

表層形　語のタイプ　左の接続情報　右の接続情報　出力形式

表層形は単語の表層形である．本辞書では，基語，接頭辞，接尾辞の他に，屈折形基語（先頭文字が脱落した形式の基語）も登録することを前提としている．また，重複語や構成語も登録することが可能である．語のタイプは，次に説明する接続テーブルの記述に用いるカテゴリであり， BASIC（基語）, PREF（接頭辞）, POST（接尾辞）などがある．左の接続情報と右の接続情報は，その語の左および右の接続に関する情報である．左の接続情報は，その形態素がどのような先頭文字を持つ形態素に接続可能かということを表す

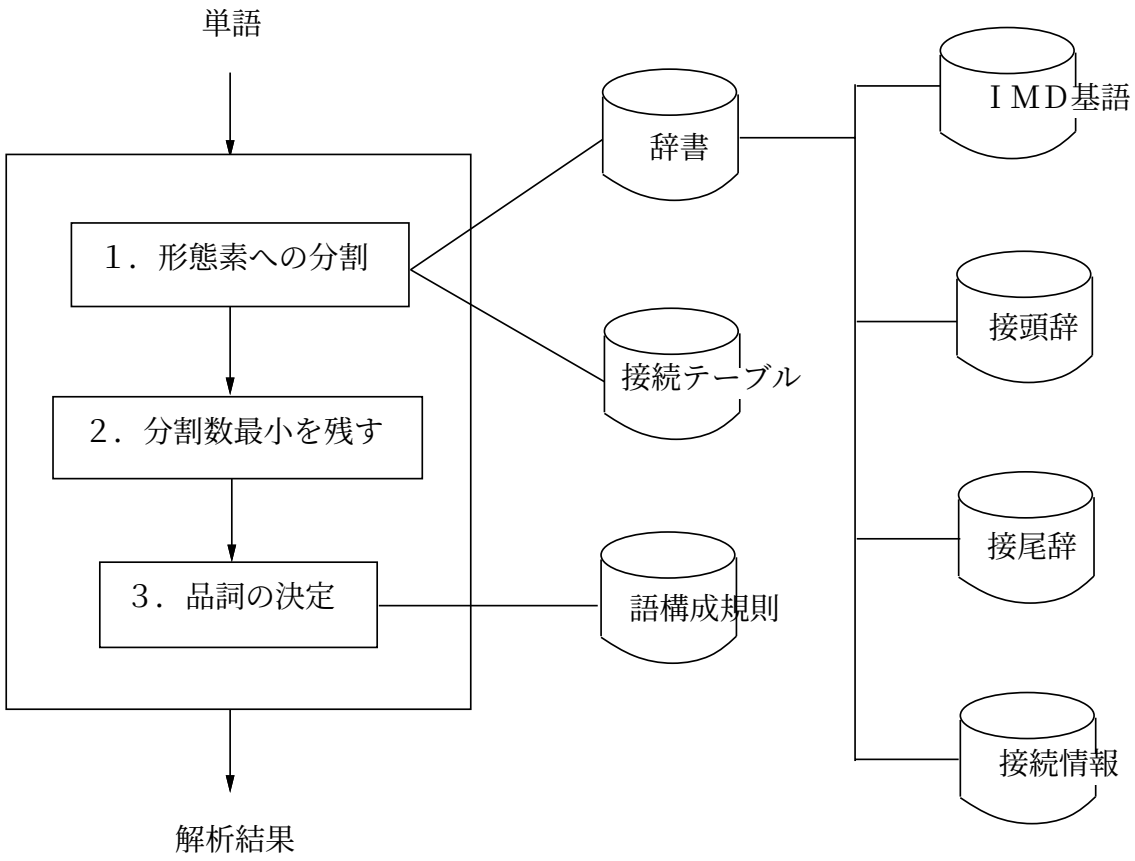


図 4.1: 形態素解析システムの構成

情報であり，右の接続情報は，その形態素がどのような先頭文字を持つかということを表す情報である。'+' は脱落しないことを意味し，'−' は脱落する（している）ことを意味する。出力形式は，その表層形に対して出力する情報であり，例えば，基語の場合は,「標準形:品詞」の形式をとる．以下に例を示す.

```
agustus     BASIC     +a     *     Agustus:NOUN
enggelam    BASIC     -t     *     tenggelam:VERB
me          PREF      +m     +l,+m,+n,+r,+w,+y     me:PREF
an          POST      +a     *     an:POST
```

具体的に使用する辞書は，前章の知識獲得の結果を用いて，ＩＭＤから次のように作成した.

- 基語は，ＩＭＤの基語を抜き出すことによって作成した．先頭の文字が k, p, s, t である基語は，先頭文字が脱落する可能性があるので，合わせて屈折形基語としても登録した.
- 接頭辞，接尾辞は，3.4節で作成したリストを登録した．このとき，3.7節で獲得した接続情報を付加した.

こうして作成した辞書のエントリーの数は，22,173 である.

4.1.2 接続テーブル

接続テーブルは，隣合う２つの形態素が接続可能であるかどうかを与えるものである．本システムの接続テーブルのフォーマットを以下に示す.

左の形態素の語タイプ　右の形態素の語タイプ　ファンクション

実際の接続テーブルの例を表 4.1に示す.

接続テーブルの１行は，２つの形態素が隣合う場合，どのような動作を行なえばよいかを示している．例えば，この表の１行目は，START という語タイプを持つ形態素が BASIC という語タイプを持つ形態素に接続するとき,「ファンクション２」を実行することによって，接続可能かどうかを調べるということを意味している．本システムがサポートしているファンクションの値は 1, 2, 3 であり，それらは，以下のことを意味している.

表 4.1: 接続テーブル

START	BASIC	2
START	PREF	2
BASIC	END	1
BASIC	HYP	1
BASIC	POST	1
HYP	BASIC	1
HYP	PREF	1
PREF	BASIC	3
PREF	PREF	3
POST	POST	1
POST	END	1
POST	HYP	1
START	COMP	2
PREF	COMP	3
COMP	POST	1
COMP	END	1

表 4.2: 語構成規則テーブル

–	an	NOUN
me	–	VERB
pe	kan	NOUN
se	an	ADVB DETR NOUN VERB

- ファンクション ＝ 1
  その２つの形態素は無条件に接続可能である．
- ファンクション ＝ 2
  形態素の左の接続情報が '–' でない限り（屈折形でない限り），接続可能である．
- ファンクション ＝ 3
  左の形態素の右の接続情報と右の形態素の左の接続情報が一致した場合のみ，接続可能である．

なお，本システムの接続テーブルに，新たにファンクションを追加することは容易に実現できる．

4.1.3 語構成規則テーブル

語構成規則テーブルは，3.6節で説明した語構成規則の集合である．語構成規則は次のようなフォーマットで記述される．

接頭辞　接尾辞　品詞の並び

具体例を表 4.2に示す．ここで，'–' は空であることを示す．

4.2 形態素解析アルゴリズム

形態素解析アルゴリズムは，前節で説明した３つの知識源に基づいて動作し，入力された単語に対して，可能な解析結果（語構成）を出力する．

本アルゴリズムは，大きく以下の２ステップに分かれる．

1. 形態素列の生成
   与えられた文字列（単語）に対して，辞書と接続テーブルを用いて，可能な形態素列を全て生成する．

2. 品詞の決定
   得られた形態素列のそれぞれに対して，語構成規則に基づいてその品詞を決定する．

4.2.1 形態素列の生成

形態素列の生成では，与えられた文字列を，まず，形態素列に分解し，得られた形態素列のうち，接続条件を満たすものだけを残すことを行なえばよい．しかしながら，それを二段階に分けて行なうのは効率が悪いので，形態素列への分解と隣合う形態素間の接続可能性チェックを同時に行なう．

形態素列生成アルゴリズム

1. 文字列 X が与えられる．

2. 形態素リスト L に'START(単語の先頭を表す)' を格納する．
   L := ('START');

3. 以下を繰り返す．

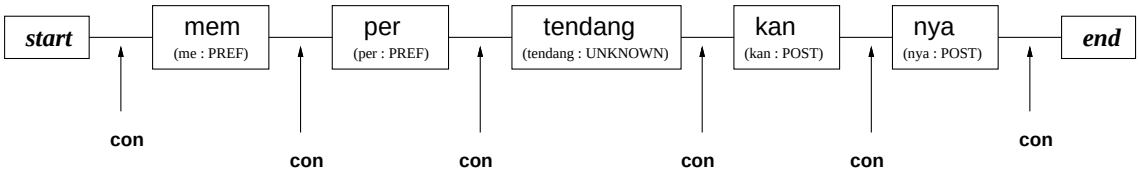
   (a) X が空ならば，L の末尾要素と'END(単語の末尾を表す)' の接続可能性を調べる．接続可能な場合は，L が求める解の一つである．これを保存し，バックトラックする．

   (b) X が空でなければ，以下を行なう．

      i. ＊文字列 X の先頭から辞書を用いて形態素を一つ切り出す．これを M とする．残った文字列を Y とする．

      ii. L の末尾要素と M の接続可能性を，接続テーブルに基づいて調べ，接続可能の場合は，M を L に追加し，Y を新たな X とする．接続不可能の場合はバックトラックする．

図 4.2: 解析例



上記のアルゴリズムで＊をつけた部分は複数の可能性が存在する部分である．本アルゴリズムは，縦型探索＋バックトラックにより，全解を探索する．

ここで，もう少し詳しく形態素への分割と接続テーブルとの関係についてに説明する．本アルゴリズムにおいて，最初の辞書引きは入力文字列の先頭で行なわれ，その直前には「START」が存在すると仮定される． START と最初に分割された入力文字列の先頭との接続可能性をチェックする．接続可能であれば，以降，辞書引きが行なわれるのは，入力文字列中において START からの有効な接続をもつ形態素の直後の位置からである．次に，その位置から残り文字列を分割し，得られた形態素とその直前の形態素の接続可能なチェックを行なう．形態素解析の最後の段階では，最後の分割された形態素と 「END 」との接続をチェックする．

mempertendangkannya の解析の様子を，図 4.2に示す．この図において， con が二つの形態素の左右の接続情報のチェックを行なうところである．

4.2.2 品詞の決定

品詞決定では，得られた形態素列のそれぞれに対して，語構成規則に基づいて品詞を決定する．そのアルゴリズムを以下に示す．

品詞決定アルゴリズム

1. 形態素列 L が与えられた．
2. L が一つの形態素から成る場合は，それは基語であり，その品詞をそのまま採用する．
3. L が複数の形態素から成る場合は，以下を行なう．
   (a) 一番先頭の接頭辞を A とする．なお，接頭辞が存在しない場合は，空とする．

(b) 一番末尾の接尾辞を Z とする．なお，接尾辞が存在しない場合は，空とする．

(c) A と Z の組で，語構成規則表テーブルを検索し，条件部が一致する規則を得る．この規則の帰結部を求める品詞とする．規則が存在しない場合は，UNKNOWN とする．

例えば， pembajakan（略奪）の場合，形態素列としては， pe:PREF ＋ bajak:NOUN ＋ an:POST が得られ，以下の規則によって，品詞が NOUN と決定される．

$$\text{pe} + X + \text{an} \rightarrow \text{NOUN}$$

また，terorisme（暴力主義）の場合，形態素列としては，teror:NOUN+is:POST+me:POST が得られ，以下の規則によって，品詞が NOUN と決定される．

$$X + \text{me} \rightarrow \text{NOUN}$$

4.2.3 分割数最小法

上記のアルゴリズムは，与えられた文字列に対して，可能な全ての解析結果（語構成）を出力する．このため，誤った解を出力することも多い．そこで，解をしぼるヒューリスティックスを導入する必要が生じる．

ここでは，分割数最小法を採用する．その場合の形態素解析アルゴリズムは以下のようになる．

1. 形態素列の生成
   与えられた文字列（単語）に対して，辞書と接続テーブルを用いて，可能な形態素列を全て生成する．

2. 得られた形態素列のうち，分割数最小のもののみを残す．

3. 品詞の決定
   得られた形態素列の品詞を，語構成規則に基づいて決定する．

なお，分割数が最小の形態素列が複数存在する場合は，本方法でも複数の解が得られることになる．

4.3 実行例

以下に本システムの形態素解析したときの実行例を示す.

1. 入力文: Peristiwa pembajakan udara itu menjadi aksi metode terorisme.
解析結果は以下に示す.

```
peristiwa       peristiwa:NOUN
pembajakan      (pe:PREF+bajak:NOUN+an:POST):NOUN
udara   udara:NOUN
itu     itu:DETR
menjadi (me:PREF+jadi:ADJV,ADVB):VERB
aksi    aksi:NOUN
metode  metode:NOUN
terorisme       (teror:NOUN+is:POST+me:POST):NOUN
```

2. 入力文: Dia menjawab semua pertanyaannya satu persatu.
解析結果は以下に示す.

```
dia     dia:PRON
menjawab        (me:PREF+jawab:UNKNOWN):VERB
semua   semua:DETR
pertanyaannya   (per:PREF+tanya:UNKNOWN+an:POST+nya:POST):UNKNOWN
satu    satu:NOUN
persatu (per:PREF+satu:NOUN):NOUN
```

第 5 章

実験と評価

本章では，本研究で提案した形態素解析システムの評価実験について述べる．

5.1 実験方法

本システムを評価するために，１００文のテキストに対して，実際に形態素解析を行ない，その結果を評価する実験を行なった．表 5.1に実験に使用したテキストの概要を示す．対象テキストは，WWW から採取した新聞記事である．

形態素解析アルゴリズムとしては，分割数最小のヒューリスティックスを組み込んだものを用いた．また，比較対象として，ＩＭＤのすべての単語を登録し，それだけを用いて形態素解析を行なう方法（すべての単語を辞書に登録しておく方法）についても，同様の

表 5.1: 実験に使用したテキスト

		数
文数		100 文
単語の延べ数		1000 語
内	基語	594 語
	構成語	406 語
単語の異なり数		703 語

表 5.2: 実験結果：得られた解の数

得られた解の数	本手法	ＩＭＤのみ
0	107 (84)	187 (159)
1	888 (317)	806 (246)
2 以上	5 (5)	7 (1)
小計 (1 以上)	893 (322)	813 (247)

実験を行なった.

5.2 実験結果と評価

入力単語に対して解析結果（語構成）がいくつ得られたかを表 5.2に示す．なお，この表において，括弧内の数字は構成語の数を表す.

この表において，得られた解の数が 0 というのは，解析不能な語，つまり，未知語となった語の数を示している．すなわち，実験に使用したテキストにおいて，ＩＭＤに存在しない単語は 187 個あったが，このうち 80 個は，ＩＭＤから獲得した知識を形態素解析で利用することにより，解析できるようになったこと．ここで，形態素解析プログラムのカバレジを

$$\text{カバレジ} \quad = \quad \frac{\text{解析できた単語数}}{\text{解析の対象となった単語数}} \tag{5.1}$$

と定義するならば，カバレジは，81.3%から 89.3%へと，8%向上したということができる.

解が得られたもののうち，どれだけ正しいかを調査した結果を，表 5.3に示す．ここで，分割が正しいとは，構成要素が正しく認識されたことを示す．分割の正解率を

$$\text{分割の正解率} \quad = \quad \frac{\text{分割が正しいもの}}{\text{解析の対象となった単語数}} \tag{5.2}$$

によって定義すると，本システムの分割の正解率は，87.6%である．これに対して，ＩＭＤのみを用いた場合の正解率は 81.3%であり，本システムはこれより 6.3%向上したと言える．なお、分割誤りは，1000 単語中 17 単語 (1.7%) であり，許容できる範囲とみなすことができる.

表 5.3: 得られた解の評価

		本手法	ＩＭＤのみ
分割が正しいもの		876 (305)	813 (247)
内	品詞が正しい	833 (268)	813 (247)
	品詞が誤り	1 (1)	0 (0)
	品詞が不明	42 (36)	0 (0)
分割が誤り		17 (17)	0 (0)
計		893 (322)	813 (247)

分割が正しいものについては，さらに，その構成語の品詞が正しく判定されているかどうかもあわせて調査した．形態素解析の真の正解は形態素の分割とその品詞がともに正しいことであるので，形態素解析の正解率は以下のように定義できる．

$$\text{形態素解析の解率} \quad = \quad \frac{\text{分割・品詞がともに正しいもの}}{\text{解析の対象となった単語数}} \tag{5.3}$$

本システムの形態素解析の正解率は 83.3%であり，ＩＭＤのみを用いた場合 (81.3%) と比較して 2.0%向上した．なお，正しく分割されたもののうち，品詞の判定を誤ったものは１語のみであり，品詞判定規則（語構成規則）は高い信頼度を持っていると考えられる．

しかし，本システムでは，品詞の決定について， 42 語についてその品詞を判定できなかった．そのうち６語は基語であり，これは，ＩＭＤにその基語の品詞が明記されていないことが原因である．残りの 36 語は構成語であり，適切な語構成規則が存在しない（ＩＭＤから獲得できなかった）ことが原因である．これらの構成語の品詞を決定するためには，さらに他の方法によって語構成規則を獲得し，本システムに追加する必要がある．

第 6 章

結論

本研究では，既存の辞書を利用して，よりカバレジの広いインドネシア語形態素解析システムを実現することを行なった．インドネシア語では，かなり複雑な語構成に関する規則を持っており，このため，インドネシア語の形態素解析の実現においては，この構成語をどう取り扱うかが問題となる．

本研究では，以下の２つのことについて研究を行なった．

- 既存の辞書からの知識の獲得
  既存の辞書に含まれる構成語を分解して，接辞や語構成規則を獲得することを行なった．

- インドネシア語形態素解析の実現
  得られた形態素知識を利用したインドネシア語の形態素解析システムを実現した．

まず，本研究では，ＣＩＣＣのプロジェクトで作成されたＩＭＤを対象とし，ＩＭＤから，接辞のリスト，接続情報および語構成規則を獲得した．

- 接辞
  形態素解析システムで使用する接頭辞３３個の内に，２１個の接頭辞を自動的に獲得することができた．システムが提案した接頭辞のうち，誤っていたものは，７個であった．一方，接尾辞については，システムが使用する４６個のうち，３９個の接尾辞を自動的に獲得することができた．システムが提案した接尾辞はすべて正しいものであった．

- 語構成規則
  ７６個の語構成規則を獲得した．その内，１４個は複数の品詞を帰結する語構成規則であった．

- 接続情報の獲得
  それぞれの接頭辞に対して，どのような文字で始まる基語に接続するかという接続情報を獲得した．

次に，これらの知識を利用するインドネシア語の形態素解析システムを作成した．本システムは，辞書，接続テーブル，語構成規則テーブル，形態素解析アルゴリズムから構成される．本システムは，まず，辞書と接続テーブルを利用して，入力単語を構成要素に分割し，次に，語構成規則を用いて，構成要素列に対する品詞を決定する．1000 単語の解析実験において，本システムのカバレジは 89.3%．正解率は 83.3%となった．これらの値はＩＭＤのみを用いた場合と比較して，それぞれ，8%，2%向上した．

本研究によって，以下のことが明らかになった．

- 既存の辞書から形態素解析に関する知識を獲得することができる．特に，語構成が複雑であるインドネシア語に対しては，このようなアプローチが有効であると考えられる．

- 既存の辞書から獲得した知識を組み込むことにより，形態素解析システムの性能を向上することができる．

しかしながら，本研究では，実際の応用に使用できる，十分なカバレジと高い正解率を持ったインドネシア語形態素解析システムを作成することはできなかった．その原因は，知識獲得の対象としたＩＭＤが十分な語彙を含んでいなかったことによる．本システムの精度をさらに向上させるためには，本システムを用いて多くのテキストを形態素解析し，その結果見つかった未知語から，本研究で提案した知識獲得の手法を用いて，新しい接辞や語構成規則を発見し，それを組み込んでいく必要がある．

謝辞

本研究を進めるにあたって，終始御指導を下さった佐藤 理史助教授に心からお礼を申し上げます.

また，終始貴重な助言を頂きました國藤 進教授，國藤研究室 助手 タナラック・ティラマヌコン博士に心から感謝致します.

最後に，本研究を色々な側面から御援助下さった，佐藤研究室，國藤研究室および奥村研究室の皆様に感謝致します.

第 A 章

付録

A.1　品詞の分類表

表 A.1，A.2，A.3，A.4に品詞の分類表を示す。

表 A.1: 品詞の分類表 (1)

Category	Code of Category	Sub Category	Code of Sub Cat
Noun	NOUN		IN
		Common Concrete Noun	INCC
		Common Abstract Noun	INCA
		Proper Noun	INPE
		Collective Noun	INPE
		Numeric Coefficient Noun	INPE
Pronun	PRON		IPN
		Personal Pronoun	IPNPS
		Indefinite Pronoun	IPNIN
		Relative Pronoun	IPNRL
		Interrogative Pronoun	IPNWH
Adjective	ADJV		IAJ
		Adjective which is gradable and can occupy both attribute and predicative in a sentence	IAJGP
		Adjective which is not gradable and can occupy both attribute and predicative in a sentence	IAJNP
		Adjective which is not gradable and can occupy the attribute position	IAJNG

表 A.2: 品詞の分類表 (2)

Category	Code of Category	Sub Category	Code of Sub Cat
Verb	VERB		IV
		Intransitive Verb	IVI
		Intransitive verb that is not need followed by a complement	IVIN
		Intransitive verb that is followed by a complement which may consist of a word, a phrase, or a clause	IVIO
		Intransitive verb that needs a prepositional phrase	IVIP
		Transitive verb	IVT
		Semi transitive verb	IVST
		Transitive verb is followed by an object	IVMT
		Transitive verb is followed by an indirect and direct object	IVBT
		Transitive verb that is followed by an object and a complement	IVCT
		Linking verb	IVLK
		Existensial verb	IVEX
Adverb	ADVB		IAD
		Adverb that modifies a verb	IADVM
		Adverb that modifies an adjective	IADAM
		Adverb that modifies a noun	IADNM
		Adverb that modifies a sentence	IADSM
Numeral	NUMR		INUM
		Cardinal Number	INUMCD
		Ordinal Number	INUMOD
		Collective Number	INUMOD

表 A.3: 品詞の分類表 (3)

Category	Code of Category	Sub Category	Code of Sub Cat
Article	ARTC		IAR
		Article that refers to non-generic concept and	IARS
		singularity. It is ussually to address a king or queen	
		Article that refers to plurality	IARP
		Article that refers to non generic concept and singularity. It may be followed by an animate or inanimate noun	IARG
Interjection	INTJ		IINJ
Phatic Category	PCAT		IPTC
Number	NUMB		INBR
Unit Symbol	USYM		IUS
Conjuction	CONJ		ICJ
		Coordinate conjunction	ICJCO
		Subordinate conjunction	ICJSO
		Correlative conjunction	ICJCR
		Extratextual conjunction which connects the new sentence or a clause with the	ICJEX
		previous one	
Preposition	PREP		IPP
Auxiliary	AUXL		IAX
		Auxiliary that has something with aspect	IAXAM
		Modal auxiliary	IAXMM
		Auxiliary that negate a verb or a noun	IAXNM

表 A.4: 品詞の分類表 (4)

Category	Code of Category	Sub Category	Code of Sub Cat
Determiner	DETR		IDET
Particle	PART		IPAR
		Interrogative marker	IPARIN
		Imperative marker	IPARIM
		Emphasis marker	IPAREM
Symbol	SYMB		ISYM

参考文献

[1] Hamman R. Yusuf, An Analysis of Indonesian Language for Interlingual Machine-Translation System, International Conference on Computational Linguistics (COLING-92), Vol.IV, pp1228-1232, 1992.

[2] Center of the International Coorperation from Computerization (CICC), Technical Report Indonesian Basic Dictionary, Machine Translation System Laboratory, CICC, 1995.

[3] 牛江清名，インドネシア語の入門，Bahasa Indonesia Untuk Para Pemula，白水社，1992.

[4] 佐藤理史, 実例に基づく翻訳, 情報処理, Vol. 33, No. 6, pp673–681, 1992.

[5] 松本裕治，今一修，山下達雄，北内啓，今村友明，日本語形態素解析システム『茶筅』 version 1.0b1 使用説明書，奈良先端科学技術大学院大学松本研究室，1996.

[6] 松本裕治, 黒橋禎夫，山地治，妙木裕，長尾真，日本語形態素解析システム JUMAN version 3.1 使用説明書, 京都大学工学部長尾研究室，1996.

[7] 渕武志，米澤明憲，日本語形態素解析システムのための形態素文法，自然言語処理，Vol.2, No.4, pp37-65, 1995.

[8] 小松英二，コスト最小法形態素解析のコストの学習法，言語処理学会第１回次大会，pp89-92, 1995.