係り受け制約の文脈自由文法への組み込み法

田辺, 利文
九州大学大学院システム情報科学研究院知能システム学専攻 : 博士後期課程

冨浦, 洋一
九州大学大学院システム情報科学研究院知能システム学専攻

日高, 達
九州大学大学院システム情報科学研究院知能システム学専攻

https://doi.org/10.15017/1474973

出版情報：九州大学大学院システム情報科学紀要. 1, pp.91-94, 1996-09-27. 九州大学大学院システム情報科学研究院
バージョン：
権利関係：

九州大学大学院
システム情報科学研究科報告
第1巻　第1号　平成8年9月

Research Reports on Information Science and
Electrical Engineering of Kyushu University
Vol.1, No.1, September 1996

係り受け制約の文脈自由文法への組み込み法

田辺利文* ・ 冨浦洋一** ・ 日高 達**

The Method of Building the Dependency Constraint into a Context Free Grammar

Toshifumi TANABE, Yoichi TOMIURA and Toru HITAKA

(Received June 24, 1996)

Abstract: In Natural Languages Processing, there are lots of syntax trees corresponding to a input sentence. If we can choose the correct syntax tree meaningfully, the quality of the processing is improved. Conventionally, we have used selectional restrictions. But semantic categories used in selectional restrictions are so rough that we couldn't choose the correct syntactic structure precisely. This paper shows the method of building the dependency constraint into a context free grammar by subdividing nonterminals according to the meaning of the phrase generated from the nonterminal and by grasping production rules from superordinate-subordinate relation of their meanings in the thesaurus.

Keywords: Context Free Grammar, The Dependency Constraint, Head word, Function, Thesaurus, Superordinate-Subordinate relation

1. はじめに

自然言語処理における構文解析では，一般に入力文に対応する構文構造がたくさん存在し，それらからどのようにして構文構造を選択するかが問題点の一つである．構文構造の中には誤った意味のものも含まれる．意味的に正しい構文構造を選択できるかどうかが，仮名漢字変換や機械翻訳などの以後の処理の質を大きく左右する．従来は文節数最少法のような粗い経験則によって構文構造の間に優先順位を付け，優先順位の高い構文構造に基づき出力を合成していた．しかし，この方法は質の高いものではなかった．そこで，さらに質を上げるには，意味処理の導入が自然言語の機械処理の大きな課題になっている．

しかし，意味処理の徹底した導入は，処理時間の爆発的な増加をもたらし，実用的ではない．従って処理の質と処理時間を考慮した意味処理の一部導入が必要である．意味処理の実用的な導入として，係り受け制約をCFG（文脈自由文法）に組み込むことが考えられる．

係り受け関係を記述することができ，確率化が容易な文法として，TAG（木接合文法）などが考えられたが[1)2)]，強力な構文解析法はまだ開発されておらず，機械処理上での重大な問題点であった．また，CFGに対しては係り受け制約を記述することが難しいとされていたが，出来ないことが証明されたわけではなかった．

本論文では，非終端記号から導出される句の概念（意味）により非終端記号を細分化し，さらにシソーラスを文法規則として捉えることにより，係り受け制約を生成規則中に取り込んだ文脈自由文法を提案する．

2. 係り受けの形式的定義

2.1 構造的な係り受けの表現

係り受け関係には，主語−述語，目的語−述語など様々な種類の関係が存在する．係り受け関係を解析する際にポイントとなるのは，係り受け関係を構成するときの係る語，係られる語と，係りの種類を決定する情報である．係りの種類を決定する情報は，日本語では格助詞や活用語尾があり，英語においては単語の位置情報や前置詞がある．

句において，修飾語，被修飾語になり得る単語をその句の*head word*，その句における係りの種類を決定する情報をその句の*function*と定義する．

今後，句Xを，非終端記号Xで用いる他に，Xから導出されている単語列としても用いることにする．

この場合，句は，

- *head word*と*function*の両方を持つ句
- *head word*を持ち，*function*を持たない句
- *function*を持ち，*head word*を持たない句

の3種類に分類することが出来る．このような句を導出する非終端記号の集合をそれぞれN_{HF}，N_H，N_Fと表すものとする．

日本語ではN_{HF}の要素(非終端記号)から導出される句としては後置詞句や動詞句などが，N_Hの要素から導出さ

平成8年6月24日受付

* 知能システム学専攻博士後期課程
** 知能システム学専攻

れる句としては名詞(句)や用言の語幹などが，またN_Fの要素から導出される句としては格助詞，副助詞や係助詞の一部などがある．

係り受けの観点から述べると，N_{HF}の要素は他の句の単語に係り得る単語とその係りの種類を決定する情報を含む句(修飾句)を導出する非終端記号であり，N_Hの要素は受けることしか出来ない単語を含む句(被修飾句)を導出する非終端記号である．N_Fの要素は単独で係りも受けも出来ない単語を含む句を導出する非終端記号であるが，その句において係りの種類を決定する単語を含む句を導出する．

自然言語の文法では，上で述べた非終端記号の分類をもとにすると，一般に次のように生成規則を分類することが出来る．

1. $X \longrightarrow Y_1 \quad \cdots \quad Y_m \quad Z$

   $Y_i(i = 1, 2, \cdots, m)$から導出される句がZから導出される句を修飾している場合．

   Y_iは係り得る句であるから$Y_i \in N_{HF}$，被修飾句は係られる句であるから$Z \in N_H$，修飾句と被修飾句が結合した句は修飾句にはなり得ず被修飾句になるので$X \in N_H$になる．Xの*head word*は，Zの*head word*になる．

2. $X \longrightarrow Y \quad Z$

   Xが修飾句になり得る場合．

   修飾句になり得るので$X \in N_{HF}$であり，修飾句は係り得る単語を含んだ句とその係りの情報(を持つ単語)を含む句で成り立つので，$Y \in N_H$，$Z \in N_F$である．Xの*head word*はYの*head word*，Xの*function*はZの*function*である．これは日本語においては従来の文節内文法に似た表現であり，すなわち自立語(句)+付属語の結合を表している．

3. $X \longrightarrow \alpha \quad X \quad \beta$

   これは単語列の結合の時にどれが*function*になるのかを規定している．$X \in N_F$である．日本語の付属語列の場合，主に係りの種類を決定するのは末尾の格助詞であり，単語列中の他の付属語は係りの種類を決定しないものとして取り扱う．英語の場合の*function*は前置詞である．この場合α，βは係りの種類を決定しない単語列である．左辺のXの*function*は右辺のXの*function*である．

文が与えられ，それに対する構文木に

$$X \longrightarrow Y_1 \quad \cdots \quad Y_m \quad Z$$

という規則(但し，$Y_i \in N_{HF}$，$Z \in N_H$，$X \in N_H$)が含まれている時，その規則においてY_iの*head word*がh_i，Zの*head word*がh，Y_iの*function*がfであるとき，h_iはhに係りの種類fで構造的に係っている(構造的な係り受け関係がある)と定義する．この係り受け関係は構文木に対して一意に決まる．

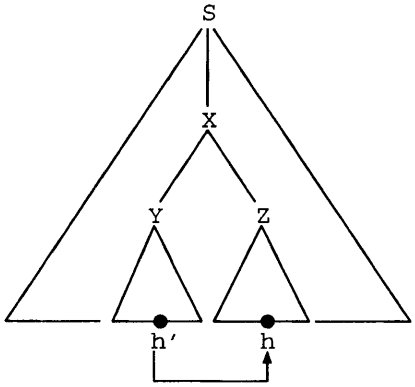

Fig.1　　h'とhが独立であること

2.2 意味的に適格な係り受けの表現

構造的な係り受け関係が意味的に適格であるとは限らない．CFGの生成規則においては構文木中の任意の規則の右辺の異なった句から導出される単語どうしは互いに独立である．

図-1は，CFGの生成規則$X \longrightarrow Y \quad Z$において，$Y$から導出される単語$h'$と$Z$から導出される単語$h$は独立であることを示している．$h'$と$h$を依存させねば係り受けは表現できない．従って，CFGで意味的にも適格に係り受け関係を表現するためには，単語と単語とを依存させることが出来るような非終端記号の細分化が必要となる．

ここで，係られる単語h_iとその係りの種類fは，係る単語hに依存すると考え，係り受けを生成規則に取り込むことを考えると，2.1の規則

$$X \longrightarrow Y_1 \quad \cdots \quad Y_m \quad Z$$

において，$Z \in N_H$から導出される語hで，$Y_i \in N_{HF}$から導出される語h_i及びf_iをコントロールすることが出来れば，意味的に適格な係り受け関係も記述することが出来る．従って，前の式は，

$$X(h) \longrightarrow Y_1(-h) \cdots Z(h) \cdots Yn(-h)$$

及び

$$Y_i(-h) \longrightarrow Y_i(h_i, f_i)$$

と2つの生成規則とすれば良い．この規則を使えば学習が比較的容易に行なえる．

この場合生成規則のパターンは次の6通りになる．

1. $Y(h, f) \longrightarrow X(h) \quad Z(f)$
2. $X(h) \longrightarrow Y_1(-h) \quad \cdots \quad X(h) \quad \cdots \quad Y_n(-h)$
3. $Y(-h) \longrightarrow Y(h', f)$
4. $Z(f) \longrightarrow \alpha \quad Z(f) \quad \beta$
5. $Y(h) \longrightarrow h$
6. $Z(f) \longrightarrow f$

ここで，次のように記法の意味を定義する．

h: *head word*

f: *function*

$X(h, f)$: *head word*がhであり，*function*がfであるカ

テゴリXの句を導出する非終端記号

$X(-h)$: *head word* がhである句に係るカテゴリXの句を導出する非終端記号

$X(h)$: *head word* がhであるカテゴリXの句を導出する非終端記号

$X(f)$: *function* がfであるカテゴリXの句を導出する非終端記号

α, β: 単語列

h及びfは，この場合単語となる．

基本的には，2.1で述べた規則を*head word*や*function*で細分化している．2番目はX から導出される*head word*(単語) h によって，Y_i から導出される句の *head word*(hに係る単語)がコントロールされることを表し，3番目はh'が係りの種類fでhに係っていることを表している．4番目は，単語列(日本語の場合は付属語列)の中のどれが*function*になるのかを規定している．*function*になり得る付属語が同一付属語列に複数含まれる場合には一般に付属語列の末尾の格助詞を*function*として用いる．格助詞がないなど複雑な場合のときは，文献[3]を参照されたい．5番目，6番目は終端記号（単語）を導出する規則である．

例えば「私がリンゴを食べる」という句を導出させるためには，次のような生成規則が必要になることがわかる．但し，開始記号はVP(食べる)とする．

VP(食べる) $\longrightarrow$ PP(−食べる) VP(食べる)
PP(−食べる) $\longrightarrow$ PP(私, が)
PP(−食べる) $\longrightarrow$ PP(リンゴ, を)
PP(私, が) $\longrightarrow$ NP(私) P(が)
PP(リンゴ, を) $\longrightarrow$ NP(リンゴ) P(を)
NP(私) $\longrightarrow$ N(私)
NP(リンゴ) $\longrightarrow$ N(リンゴ)
VP(食べる) $\longrightarrow$ V(食べる)
N(私) $\longrightarrow$ 私
N(リンゴ) $\longrightarrow$ リンゴ
P(が) $\longrightarrow$ が
P(を) $\longrightarrow$ を
V(食べる) $\longrightarrow$ 食べる

尚，今回提案した文法では，以下を取り扱わない．

- 非交差性を満足しない文：係り受けの重要な性質として，係り受けの非交差性があるが，それを満たさない言語が少数であるが存在し，日本語においてもそのような文も見受けられる．しかしそのような文は非常に少ないものとして，ここではそのような文を除外して考える．CFGの範疇内では，係り受けに交差を有する文の構文解析は出来ない．
- 並列句を含む文：並列句の場合は生成規則の右辺に*head word*が複数あるため左辺にも複数の*head word*が必要になり，可算無限個の非終端記号が存在してしまいCFGの範疇から外れるため除外する．
- 終助詞：*head word*にはならず，*function*としても機能しないので，取り扱わなくても支障がない．
- 受身，使役の助動詞：これらは本来は*function*であり，文全体の格関係を大きく変えてしまう．これをどう取り扱うかはこれからの研究課題とする．

従来のCFGによる構文解析においては選択制約を意味処理の過程で行なっていた．選択制約とは，簡単にいえば，意味的に適格な係り受け関係であるかを調べ，適格ではないものを排除することである．上の記法では，この選択制約をも生成規則として記述出来ることを意味する．従って，提案した文法を確率化することにより，選択制約をも考慮した構文木の生起確率を簡単に決定することが出来る．例えば，一つの文に対する構文木の順位つけを行なう場合，PCFG による構文木の確率が高いが意味的に適格な係り受け関係があまり成立していないものと，構文木の確率は低いが意味的に適格な係り受け関係が成立している場合，このどちらの構文木の順位を上げるかが問題になってくる．上記の記法では，そのことを全く意識しなくて済む．

3. 生成規則へのシソーラスの組込み

2.2節で提案した文法において問題になるのは，生成規則の数である．*function*として考えられるものについては格助詞や活用の種類ぐらいしかないが，*head word*として考えられるものについては，名詞の他に動詞などがあり，その数も膨大になる．2.2の3番目のタイプの形の生成規則は，2つの語(*head word*)の間に係り受け関係があることを意味している．そこでは，*head word*を単語としている．しかしその場合，非終端記号の数が膨大になり，その結果生成規則の数が増え，処理時間も増えてしまう．従って，意味的に適格な係り受け関係の表現精度にあまり影響を及ぼさないように非終端記号を減らす方法を考える必要がある．

3.1 シソーラス

単語Aが指示する対象の性質を単語Bの指示する対象も持つ時，単語AとBには上位下位関係が成立し，単語Aは Bの上位語であるという．

シソーラスは概念間の上位下位関係を体系的に表したものである．図-2はその例である．

上位下位関係では推移律が成り立つ．例えば，「生物」が「動物」の上位語で，かつ「動物」が「人間」の上位語であるということから，「生物」が「人間」の上位語であることが分かる．

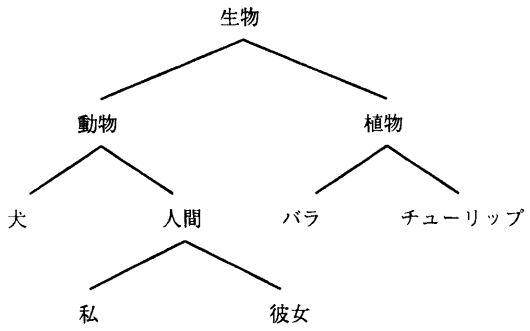

Fig.2　名詞のシソーラスの例

3.2 生成規則への組込み

単語w_1とw_2の間に意味的に適格な係り受け関係が成立していれば，w_1の下位語であるw_1'と，w_2の下位語であるw_2'の間にも意味的に適格な係り受け関係が成立するものと仮定する．

シソーラスにおいて概念 W_u と W_d が上位-下位関係であるとき，

$$W_u \longrightarrow W_d$$

の生成規則として捉え，単語 w の概念が W であるとき，

$$W \longrightarrow w$$

の生成規則として捉える．

従って，シソーラスを用いた場合の文法の生成規則のパターンは次の8通りとなる．

1. $Y(H,f) \longrightarrow X(H) \quad Z(f)$
2. $X(H) \longrightarrow Y_1(-H) \quad \cdots \quad X(H) \quad \cdots \quad Y_n(-H)$
3. $Y(-H) \longrightarrow Y(H',f)$
4. $Z(f) \longrightarrow \alpha \quad Z(f) \quad \beta$
5. $Z(f) \longrightarrow f$
6. $Y(H) \longrightarrow H$
7. $H \longrightarrow H'$
8. $H \longrightarrow h$

但し，記号Hはシソーラス中のある概念記号であり，これを*head word* として用いている．h，fは単語である．6番目のタイプの生成規則は，非終端記号からシソーラスの概念記号への書換えを行なっている．7番目はシソーラスの概念記号をたどっている．8番目はシソーラスの概念記号から単語への書換えを行なっている．

シソーラスを用いない従来の文法をG_1，シソーラスを代わりに用いた場合の文法をG_2とすると，*head word*を適当に選ぶことにより，

$$L(G_1) \simeq L(G_2)$$

とすることが出来る．シソーラスの比較的上位の概念記号を*head word* とすると，生成規則は減るが意味的に適格な係り受け制約が粗くなる．逆にシソーラスの下位の概念記号を*head word*とすると，意味的に適格な係り受け関係は細かくなるが生成規則の数が膨大になる．そのため，これらを考慮した*head word*の選定が問題となる．

4. おわりに

文脈自由文法の非終端記号をそれから導出される句の有限個の概念（*head word*）及び*function*により細分化することで，これまで文脈自由文法の枠組では表現できないとされていた係り受け制約を表現できる文脈自由文法の構成法を提案した．

構文的制約及び係り受け制約をともに満たす解析結果（構文木）が複数ある場合に，構文木の尤もらしさを基に構文木に優先順位を付けることが考えられる．文脈自由文法を確率化した確率文脈自由文法では，この尤もらしさを構文木の生起確率として与えることが出来る．確率文脈自由文法は，統計モデルの一つとして考えられ，このモデルのパラメータは事例データ（標本）から最尤推定により求めることができ，こうして求めた確率文脈自由文法は，事例の統計的性質を反映したものとなる．従って，係り受け制約を生成規則として取り込んだ文脈自由文法を確率文法化することにより，事例の係り受けに関する統計的な性質を考慮した優先づけが出来る．

*head word*としてシソーラス中の概念を用い，概念間の上位下位関係を生成規則として捉えた．これにより意味的な正確さと処理時間のトレードオフに対し，シソーラス中の任意の概念を*head word*として選ぶことが出来るので，それぞれの自然言語処理システムの要求するレベルに合わせることが出来る．

今後は，*head word*をどの概念にするのが妥当であるか，*function*をどういう風に決めるかを検討する．

参考文献

1) Joshi, Aravind K., Yves Schabes: Tree Adjoining grammars and lexicalized grammars, In Maurice Nivat and Andreas Podellski, editors, Tree Automata and Languages, Elsevier Science 1992
2) Yves Schabes, Richard C.Waters: Lexicalized Context-Free Grammars, In proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, 1993
3) 渡辺健一郎: 付属語列文法に対する一考察, 九州大学工学部学士論文, 1996