2006－NL－171（11）
2006／1／13

生成のための日本語LFG文法の構築

大熊　智子† 　増市　博† 　吉岡　健†

† 富士ゼロックス (株) 研究本部
〒 259-0157 足柄上郡中井町境 430 グリーンテクなかい
E-mail: †{Ohkuma.Tomoko,Hiroshi.Masuichi,Yoshioka.Takeshi}@fujixerox.co.jp

あらまし

Lexical Functional Grammar(LFG)，Head Phrase Structure Grammar(HPSG) などの単一化文法を用いたパーザは，入力された自然言語文に対し，functional-structure や minimal recursion semantics(MRS) などの統語意味構造を出力する．このプロセスを逆に辿ることによって，統語意味構造を入力とし，同じ文法を用いて自然言語文を出力として得ることができる．すなわち，パーザに用いられた文法をそのままジェネレータに適用することが可能である．LFG に基づく処理系である Xerox Linguistic Enviroment(XLE) や HPSG の代表的な処理系である LKB も，ジェネレータ機能を有している．このような生成技術の代表的な応用先として，中間言語方式の翻訳システムを挙げることができる．また，最近注目を集めている「言い換え」に対しても生成技術を適用可能である．さらに，QA システムや対話システムなど，生成技術は様々な言語処理アプリケーションに適用可能な基礎技術である．しかしながら，実際に解析用の日本語文法をそのまま生成に適用しようとすると，解析の段階では問題にならなかった様々な課題が顕在化した．本報告書では，我々が研究開発を進めてきた解析用文法を用いて生成を行う際の課題を分析し，それを解決するための手法を提案した．さらに文の生成実験を行い，それらの妥当性を検証した．実験結果から，例外ルールの付加と語彙の選択という二種の手法が生成成功率の向上に寄与することを確認できた．

キーワード　生成 語彙機能文法 LFG 日本語文法 曖昧性

Adapting Japanese LFG grammar to generation

Tomoko OHKUMA†, Hiroshi MASUICHI†, and Yoshioka TAKESHI†

† Corporate Research Center Fuji Xerox Co., Ltd.
430 Sakai, Nakai-machi, Ashigarakami-gun, Kanagawa,Japan
E-mail: †{Ohkuma.Tomoko,Hiroshi.Masuichi,Yoshioka.Takeshi}@fujixerox.co.jp

Abstract

A parser based on unification grammars such as Lexical Functional Grammar (LFG) and Head-driven Phrase Structure Grammar (HPSG) outputs functional structures and MRS's for an input sentence.A generator takes the inverse process of a parser. For example, the parser functionality of Xerox Linguistic Environment (XLE) accepts a sentence as input and produces f-structures as output; on the contrary, the generator functionality of XLE accepts an f-structure as input and produces all the sentences that, when parsed, could have the input f-structure as output.The generator can be useful as a component of translation, paraphrase, question answering system, and a natural language interface. However problems appeared when we applied the Japanese LFG grammar for a parser to the XLE generator. We examined the problems and proposed techniques for lexical rules and grammar rules to solve the problems. Then we evaluated the techniques conducting experiments. The results of the experiments show that the techniques we proposed can improve the coverage of the generation functionality of XLE.

Key words　Generation, Lexical Functional Grammar, LFG, Japanese grammar rules, ambiguity

1. はじめに

1.1 生成とは

Lexical Functional Grammar(LFG), Head Phrase Structure Grammar(HPSG) などの単一化文法に基づくパーザでは，入力された自然言語文に対し，f(unctional)-structure や minimal recursion semantics(MRS) などの統語意味構造を出力する．このプロセスを逆に辿ることによって，統語意味構造を入力として同じ文法を用いて自然言語文を出力することができる (Kay 1996)．すなわち，パーザに用いられた文法をそのままジェネレータに適用することが可能である．LFG に基づく処理系である Xerox Linguistic Enviroment(XLE) や HPSG の代表的な処理系 LKB も，ジェネレータ機能を有している (Maxwell and Kaplan 1993)(Caroll and Oepen 2005)．

このような生成技術の代表的な応用先として，中間木を利用した翻訳システムを挙げることができる．(Frank 1999) では，f-structure を中間木にした英仏翻訳について提案している．また，MRS を中間木として利用した翻訳システムの研究もされてきた (Copestake, Flickinger, Malouf, Riehemann, and Sag 1996)．最近では，LFG と HPSG の両方のシステムを用いたノルウェー語–英語間の翻訳を目指す LOGON プロジェクトの活動も紹介されている (Oepen, Dyvik, Flickinger, Lonning, Meurer, and Rosen. 2005)．

さらに，(乾 藤田 2004) などで紹介されている「言い換え」に対しても生成技術を適用可能である．f-structure の操作を行えば，その結果生成される文を「言い換え」ることができるし，そもそも特別な操作を行わなくても語順の入れ換えなどが生じれば，結果的に「言い換え」が行われたとみなすことができる．これ以外にも，QA システムや対話システムなど，生成技術は様々な言語処理アプリケーションに適用可能な基礎技術である．

1.2 本研究の目的

実際に解析用の日本語文法をそのまま生成に適用しようとすると，解析の段階では問題にならなかった事柄が顕在化する．本研究では，XLE 上で動作する解析のための大規模 LFG 日本語文法 (増市 大熊 2003) を用いて生成を行う際の課題を分析し，それを解決するための手段について提案する．さらに文の生成実験を行ってそれらの妥当性について検証する．本研究の最終的な目的は，生成のための大規模日本語文法を構築することである．

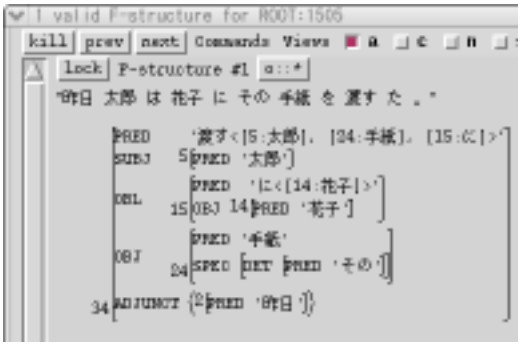

図 1　入力となる f-structure の例

本稿の構成は以下の通りである．2 章では，解析のための文法を生成に適用する際に顕在化する問題点について述べる．3 章では，2 章で述べた問題点を解決するための手当てについて説明し，4 章でこの手法の妥当性を検証する実験結果を示し，考察を行う．5 章でまとめと今後の課題について述べる．

2. 生成を目的とした日本語文法の課題

2.1 解析用文法を適用して文の生成を行う際の問題点

連用修飾成分の語順が自由であり，述語に含まれる助詞，助動詞の連接パターンが多岐にわたる日本語にとって，文生成の際に最も大きな障壁となるのは，生成結果の曖昧性の爆発である．限られた言語事象を対象とした TOY レベルのパーザでは問題にならないが，実用を視野に入れた大規模文法は，多数のルールを含むため，本来生成されることを意図していなかった文が多く生成される可能性が高い．図 1 に示した f-structure を入力として生成された 24 文を下記に示す．XLE では生成結果はすべて圧縮された形式で出力される．このように，単純な構造を持つ f-structure からでも，複数の文が生成される．

{ { {昨日 花子 に|花子 に 昨日} その 手紙 を
|その 手紙 を {昨日 花子 に|花子 に 昨

日}
|昨日 その 手紙 を 花子 に
|花子 に その 手紙 を 昨日}
太郎 は
|太郎 は
{ {昨日 花子 に|花子 に 昨日} その 手紙 を
|昨日 その 手紙 を 花子 に
|花子 に その 手紙 を 昨日
|その 手紙 を {昨日 花子 に|花子 に 昨日}}
| { {昨日 花子 に|花子 に 昨日} 太郎 は
|昨日 太郎 は 花子 に
|花子 に 太郎 は 昨日}
その 手紙 を
|その 手紙 を
{ 太郎 は {昨日 花子 に|花子 に 昨日}
|昨日 太郎 は 花子 に
|花子 に 太郎 は 昨日}
|昨日 {その 手紙 を 太郎 は|太郎 は その 手紙 を} 花子 に
|花子 に {その 手紙 を 太郎 は|太郎 は その 手紙 を} 昨日}
渡す た .

上記の 24 通りの結果はすべて連用修飾成分の語順によるものであり，その全てを正解と言うことができる．しかし，一般には生成された複数の文の中には非文が含まれる場合もある．非文を生み出す原因は，文法ルールと語彙ルールの両方に存在する．下にそれぞれの問題点について述べる．

2.2 文法ルールにおける問題点

大規模文法は，様々な形態の文を解析するために，比較的緩やかな制約で記述されている．この緩やかな制約が，非文を生む要因となる．

例えば，下記の (1) のように動詞句の中で動詞に連接する n 個の助詞 AUX のあらゆる組み合わせを受けつけるルールは，文を無限に導出しようとすることになり，その結果引き起こされるメモリ不足が文の生成を阻む．

(1) $\mathrm{VP} \longrightarrow \underset{(\uparrow=\downarrow)}{\mathrm{V}} \{AUX_1|AUX_2|AUX_3|...AUX_n\}^*.$

このルールのように，非文の排除を念頭に置かずに記述した文法は非文を含む多くの文を生成してしまう．

一方で，制約を厳しくすればするほど，解析カバー率は低下する．(鳥澤 1999) でも指摘されているように，日本語のすべての言語現象を網羅的に記述することは困難である．言語事象を一つ一つ個別に羅列するのではなく，上記のルールのようなかたちで，ある程度の抽象化を行わなければ，高い解析カバー率は達成できない．

つまり，解析カバー率の向上と過生成の低減はトレードオフの関係にあると言える．様々な構文を受け付けるために，文法ルールの抽象度を高めれば高めるほど，そのルールは非文を含む多数の文を生成することになるし，文の生成を抑えるために各言語事象のためのルールを個別に書いて抽象度を下げれば，解析できない文を増やすことになるからである．

2.3 語彙ルールの問題点

語彙ルールで定義される Lexical Category(LC) には複数の単語が含まれる.「かも知れない」,「かも知れぬ」のようなパラフレーズ群には同一の LC を与えている．これらに別々の LC を与えておくことは，結局，すべての表層に別々の LC を与えることになり，語彙ルールを記述することは事実上不可能となってしまう．解析の際には問題にならないが，生成を行う際にはこの同一の LC に含まれる語は出力結果数を増大させる．LC に含まれる語の数だけ，生成される文のバリエーションが存在するからである．

それでも PRED やなんらかの属性を持つ LC はその情報が，f-structure に含まれている以上，たとえ属する LC が同じでも，他のメンバとの区別ができるので，過生成の原因にはなりにくい．しかし，PRED や属性を持たない LC の場合，f-structure の中に情報が残されず，もしその LC が文法ルールに含まれる場合はすべての LC が生成結果に含まれてしまうため，過生成の大きな原因となりうる．現在の解析用語彙ルール中で助詞に相当する LC には，PRED も属性も持たないものが存在する．例えば，一部の終助詞がこれにあたる．

3. 過生成を防ぐための処理

3.1 過生成制御の方針

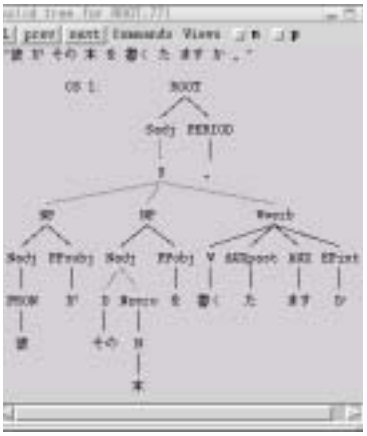

図 2　非文を導出する c-structure の例

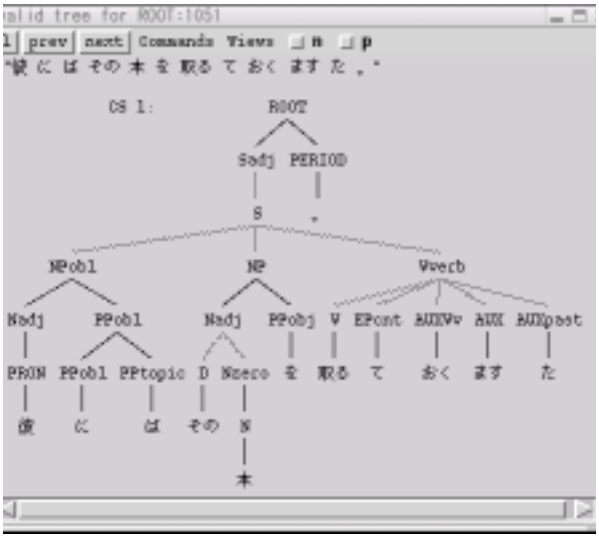

図 3　複合動詞を含む文の c-structure

生成の目的，つまりどのようなアプリケーションに適用するかによって，解を絞り込む方針は異なる．例えば，冒頭で述べた機械翻訳に生成を適用する場合，最終的に得たい日本語文は１つに絞られなければならない．

ところが，(乾・藤田 2004) で紹介されているように，情報検索やＱＡシステムなどでクエリー拡張のための言い換えを実施するための手段として用いることを想定した場合，むしろ結果は絞られるべきでなく，それが日本語として正しい文である限り，なるべく多くの

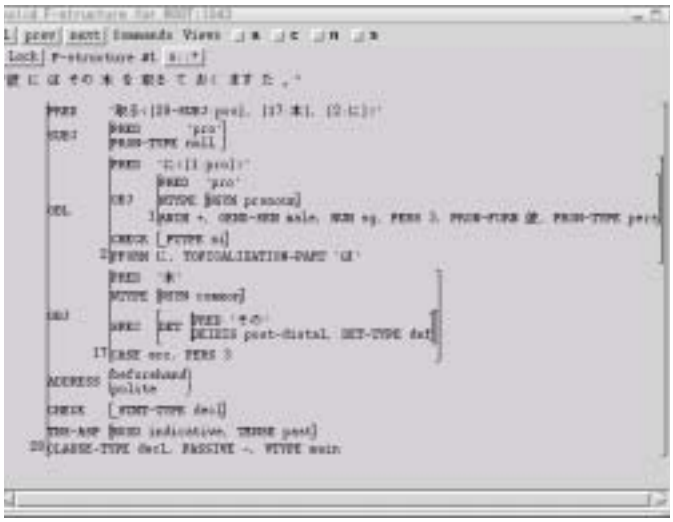

図 4　複合動詞を含む文の f-structure

文が得られる方が望ましい．つまり，2.1 節で示した生成結果から唯一の解を得ようとする処理は必要ない．

以上のことを踏まえて，本稿では連用修飾の語順に制限を加えず，非文を生み出す原因となりやすい述部の語彙と文法に着目して，処理を行った．

3.2 例外ルールの付与

生成結果に含まれる非文とそれを導出した図 2 のような c(onstraint)-strcture，つまり構文木を見て，どのようなルールが非文を生成しているのか観察する．一つの語は二つ以上の LC に属している場合もあるため，この作業は単純に表層あるいは形態素の並びではなく，c-structure を対象に行わなくてはならない．そして，この観察によって確認された非文を生成するルールを，既存の解析ルールに例外ルールとして付与する．

下記に 2.1 節に示した (1) に例外ルールを付与した例を (2) に示す．この例は，AUX_3，AUX_1 の並びが，動詞の直ぐ右には現れないことを表現している．

$$(2)\quad \underset{(\uparrow=\downarrow)}{\mathrm{VP}\longrightarrow\ \mathrm{V}}\quad \underset{-AUX_3AUX_1}{\{AUX_1|AUX_2|AUX_3|...AUX_n\}^*}.$$

この方式の利点は主に二つある．まず一つは，作業効率である．解析のためのルール記述が，日本語として正しい文を受け入れる規則の発見に重点が置かれるものであるのに比べ，生成のための記述はむしろ非文を排除することが重要になってくる．したがって,「ありえ

表 1　EDR コーパスから抽出した 5,000 文の解析結果

文の数 (%)	平均語数	正常解析 (文) (%)	部分解析 (文) (%)	解析失敗 (文) (%)	解の平均数 (個)
5,000 (100%)	21.3	4356 (87.1%)	459 (9.2%)	185 (3.7%)	16.0

表 2　各条件と生成の成功率

	OT マーク	
例外条件	OT_all	OT_select
有	88.22%	95.68%
無	89.47%	95.06%

ない」構文規則を例外として記述していくことは，生成のための文法構築において直感的かつ効率的な作業であるといえる．

もう一つは，解析カバー率の保持である．文法ルールの実装における現実問題として，ルールが複雑になりすぎて編集が事実上不可能な場合もある．こういった場合に，一度構築した文法ルールを修正して，制約の強い文法に変更することは非常に困難で危険である．この編集によって，今まで解析できていた構文が解析できなくなる可能性があるからである．

3.3　語彙の選別

日本語 LFG 文法では，各機能的注釈に対して，Optimality Theory (Bresnan 2001) に基づいたマーク (OT マーク) を付与している．OT マークには予め優先順位を設定しておく．優先順位の高い OT マークが付与された機能的注釈に基づいて得られた f-structure を優先的に最終結果に残す．OT マークを付与する本来の目的は言語学的な根拠に基づいて解析結果の曖昧性を減少させることである．我々もこの目的のために OT マークを使用し，解析結果数の爆発を防いできた (増市・大熊 2003)．ここでは，過生成の大きな要因である PRED と属性を持たない LC に優先順位を最低にする OT マークを付加する．ただし，文の生成の際にのみ働く生成用 OT マークを付与し，生成の際に LC が無視されるようにする．

なお，文法ルールの中で生成の際に重要な役割を果たす LC で，PRED および属性を持たないものも存在する．例えば，複合動詞を形成する助詞「テ」などがこれに当たる．図 3 にこのタイプの複合動詞を含む c-structure を図 4 に f-structure を示す．f-structure の中には「テ」に関する情報は含まれていないが，c-structure では二つの動詞の結合子として，解析，生成の際に非常に重要な LC として存在しているのが分かる．そこで，終助詞のように構文の形成にそれほど寄与しない LC と，このように構文の形成に必要不可欠な LC を分けた．

4. 実　　験

4.1　手　続　き

(増市・大熊 2003) の日本語 LFG 文法を用いて，EDR コーパスからランダムに選んだ 5,000 文を解析した．表 1 に解析結果を示す．

上記の解析で一つ以上の解が得られた 4,815 文それぞれに対して一つの f-structure を得る．複数の f-structure が存在する場合には，ランダムに一つを選択する．得られた f-structure を入力として文を生成する．一つ以上の文が生成できれば生成成功とする．OT マークの各条件は以下のように設定した．

- OT_all: PRED と feature を全く持たない LC に OT マークを付与
- OT_select: OT_all で OT マークを付与した LC のうち，重要な LC(3.3 節参照) から OT マークを外す

この二つの条件のもとで，例外ルールを付与した場合としない場合を比較する．

4.2　実験結果と考察

表 2 に実験結果を示す．例外ルールの有無によって，生成成功率の結果に大きな差はみられなかった．ただし，速度向上には非常に効果的であった．現在の例外ルールは 5 個程度と少なく，さらに多くの例外ルールを追加していく予定である．

LC への OT マークの付加は生成成功率を左右することが分かった．ただし，やみくもに付与すると，文法ルールに影響を及ぼし，その結果生成に失敗する可能

性もあるため，重要な LC を見極めることが今後の課題である．

5. おわりに

本研究では，例外ルールの付加と語彙の選択という二つの手法を用いて，解析用文法を生成に適応させる手法を提案し，両者が生成成功率の向上に寄与することを確認した．

今後は精度の測定を行うことによって，非文の排除のために，例外ルールの充実化をはかる．

ただし，解析と比べて生成のための正解コーパスを作成することは難しい．何故なら，解析結果よりも生成結果の方がより多くの複数解を持つため，人手であってもすべての解を網羅的に記述することは容易ではないためである．したがって，精度の測定手法の確立自体も今後の課題の一つである．

また，今回行った例外ルールの付加と語彙への OT マークの付加はいずれも，文法記述者の主観によって行った．しかし，精度の測定を行う手法が確立できれば，統計的な手法によって両者を拡張することも可能であると思われる．

参考文献

Caroll, J. and Oepen, S. (2005). "High Efficiency Realization for a Wide-Coverage Unification Grammar." *Procceedings of IJCNLP-05*, 165–176.

Copestake, A., Flickinger, D., Malouf, R., Riehemann, S., and Sag, I. (1996). "Translation Using Minimal Recursion Semantics." *Proceedings of The sixth International Conference on Theoretical and Methodological Issues in Machine Translation.*

Frank, A. (1999). "From Parallel Grammar Development towards Machine Translation." *Procceedings of MT Summit VII.*

乾健太郎 藤田篤 (2004). "言い換え技術に関する研究動向." 自然言語処理, **11** (5), 151–198.

Bresnan, J. (2001). *Optimal syntax*, pp. 334–385. Oxford University Press.

Kay, M. (1996). "Chart Generation." *Procceedings of the 34th Meeting of the Association for Computational Linguistics*, 200–204.

増市博 大熊智子 (2003). "日本語 LFG による大規模構文意味解析システムの構築." 自然言語処理, **10** (2), 79–109.

Maxwell, J. T, I. and Kaplan, R. (1993). "The interface between phrasal and functional constraints." *Computational Linguistics*, **19**, 571–589.

Oepen, S., Dyvik, H., Flickinger, D., Lonning, J. T., Meurer, P., and Rosen., V. (2005). "Holistic regression testing for high-quality MT. Some methodological and technological reflections." *Proceedings of the 10th Annual Conference of the European Association for Machine Translation.*

鳥澤健太郎 (1999). "高機能な構文解析器に向けて–HPSG のための実用的な構文解析器–." 情報処理, **40** (4), 380–386.