日本語 LFG にもとづく助数詞の処理

大熊 智子[†]・梅基 宏[†]・三浦 康秀[†]・増市 博[†]

事物の数量的側面を表現するとき、数詞の後に接続する語を一般に助数詞と呼ぶ。英語などでは名詞に直接数詞が係って名詞の数が表現されるが、日本語では数詞だけでなく助数詞も併せて用いなければならない。名詞と助数詞の関係を正しく解析するためには、助数詞が本来持つ語彙としての性質と構文中に現れる際の文法的な性質について考慮する必要がある。本稿では、数詞と助数詞の構文を解析するための Lexical-Functional Grammar (LFG) の語彙規則と文法規則を提案し、その規則の妥当性と解析能力について検証した。提案した規則によって導出される解析結果 (f-structure) と英語、中国語の f-structure をそれぞれ比較することによって、日本語内での整合性と多言語間との整合性を有していることが確認できた。また、精度評価実験の結果、従来の LFG 規則に比べて通貨・単位に関する表現では 25%、数量に関する表現では 5%、順序に関する表現では 21% の F 値の向上が認められた。

キーワード：助数詞、語彙機能文法、文法規則、構文解析

Treatment of Numerical Classifier Based on Lexical-Functional Grammar

TOMOKO OHKUMA[†], HIROSHI UMEMOTO[†], YASUhide MIURA[†] and HIROSHI MASUICHI[†]

Japanese numeral classifier (NC) expresses quantificational expressions by following number nouns. The parser based on phrase structure grammar like Lexical-Functional Grammar (LFG) or Head-driven Phrase Structure Grammar (HPSG) has to define the relationships between noun and NC or number and NC by the grammatical rules. NC has various relationships syntactically because of its' various characteristics. The treatment of NCs in LFG formalism should take account of their lexical and syntactical characteristics. This report proposed LFG rules for NCs and examined their validity and accuracy. The comparison between Japanese f-structure inducted by the rules proposed in this report and English f-structure and Chinese f-structure which correspond to the Japanese f-structures show that the rules make f-structures parallel between same phrases in Japanese and in other languages. The experiment of evaluations shows the rules improve F-score of NCs for currency and unit from 52.10% to 77.90% and F-score of NCs for quantity from 81.92% to 87.32% and F-score of NCs for Order from 59.62% to 81.26%.

Key Words: *Numerical Classifier, Lexical-Functional Grammar, grammar rules, syntactic parsing*

[†] 富士ゼロックス（株）研究技術開発本部システム要素技術研究所, Research and Technology Group, System Technology Laboratory, Fuji Xerox Co., Ltd.

1 はじめに

1.1 背景

事物の数量的側面を表現するとき, 「三人」, 「5 個」, 「八つ」のように, 「人」, 「個」, 「つ」という付属語を数詞の後に接続する. これらの語を一般に助数詞と呼ぶ. 英語などでは “3 students”, “5 oranges” のように名詞に直接数詞が係って名詞の数が表現されるが, 日本語では「3 人の学生」, 「みかん五個」のように数詞だけでなく助数詞も併せて用いなければならない.

形態的には助数詞はすべて自律的な名詞である数詞に付属する接尾語とされる. しかし, 助数詞の性質は多様であり, 一律に扱ってしまうことは統語意味の見地からも計算機による処理においても問題がある. また構文中の出現位置や統語構造によって, 接続する数詞との関係は異なる. つまり, 数詞と助数詞の関係を正しく解析するためには 1) 助数詞が本来持つ語彙としての性質, そして 2) 構文中に現れる際の文法的な性質について考慮する必要がある.

KNP (黒橋 2000) や cabocha (工藤, 松本 2002) などを代表とする文節単位の係り受け解析では, 上記のような数詞と助数詞の関係は同じ文節内に含まれるため, 両者の関係は係り受け解析の対象にならない. ところが, 単なる係り受け以上の解析, 例えば Lexical Functional Grammar (以下, LFG) や Head-driven Phrase Structure Grammar (以下, HPSG) のような句構造文法による解析では, 主辞の文法的役割を規程する必要がある. つまり文節よりも細かい単位を対象に解析を行うため, 名詞と助数詞の関係や数詞と助数詞の関係をきちんと定義しなければならない.

上記のような解析システムだけでなく, 解析結果を用いたアプリケーションにおいても助数詞の処理は重要である. (梅基, 杉原, 大熊, 増市 2008) で紹介されている検索システムにおける含意関係の判定では数量, 価格, 順番などを正しく扱うことが必要とされる.

1.2 本研究の目的

本稿では, 数詞と助数詞によって表現される構文¹を解析する LFG の語彙規則と文法規則を提案し, 計算機上で実装することによってその規則の妥当性と解析能力について検証する. これらの LFG 規則によって出力された解析結果 (f-structure) の妥当性については, 下記の二つの基準を設ける.

(1) 他表現との整合

統語的に同一の構造を持つ別の表現と比較して, f-structure が同じ構造になっている.

(2) 他言語との整合

他の言語において同じ表現の f-structure が同じ構造になっている.

¹ 但し, 「3 年」, 「17 時」など日付や時間に関する表現は (Bender and Siegel 2004) と同様にこの対象範囲から除く.

2 章では助数詞に関する従来研究を概観し、特に関連のある研究と本稿の差異について述べる。3 章では助数詞のための LFG 語彙規則と助数詞や数詞を解析するための LFG 文法規則を提案する。4 章では 3 章で提案した LFG 規則を (増市, 大熊 2003) の日本語 LFG システム上で実装し、システムによって出力される f-structure の妥当性を上記の二つの基準に照らして検証する。日本語と同様にベトナム語や韓国にも日本語のそれとは違う性質をもった固有の数詞と助数詞が存在する (矢崎 1999)。また、日本語の助数詞は一部の語源が中国語にあるという説もあり、その共通性と差異が (渡辺 1952) などで論じられている。そこで、Parallel Grammar Project (Butt, Dyvik, King, Masuichi, and Rohrer 2002) (以下, ParGram) において LFG 文法を研究開発している中国語 LFG 文法 (Fang and King 2007) で導出された f-structure を対象にして、基準 2 を満たしているかを確認するために比較を行う。

“3 kg” の ‘kg’ や, “10 dollars” の ‘dollar’ など、英語にも数字の後に接続する日本語の助数詞相当の語が存在する。また、日本語においても英語のように助数詞なしに数詞が直接接続して名詞の数量を表現する場合もある。ParGram において英語は最初に開発された LFG 文法であり、その性能は極めて高い (Riezler, King, Kaplan, Crouch, Maxwell, and Johnson 2002)。ParGram に参加する他の言語は必ず英語の f-structure との比較を行いながら研究を進める。以上のことから、中国語だけではなく (Riezler et al. 2002) の英語 LFG システムで出力された f-structure との比較を行う。5 章では精度評価実験を行って、解析性能を検証する。数詞と助数詞によって形成される統語を LFG 理論の枠組みで解析し、適切な f-structure を得ることが本研究の目的である。

2 先行研究

国語学や言語学の分野において、助数詞に関する研究は古くから行われてきた。また、文法的な性質だけでなく、助数詞や助数詞の数える対象である名詞との関係に着目して、意味的な分類を行う試みもある。さらに、自然言語処理の分野では助数詞の特徴を考慮して処理を行う解析手法が提案されている。本章では助数詞の文法に関する研究、助数詞の分類に関する研究、助数詞を対象にした自然言語処理についてそれぞれ概観する。

2.1 助数詞の文法に関する研究

日本語助数詞の代表的な研究に (見坊 1965) が挙げられる。形態的には「最も単純な品詞」とされてきた助数詞であるが用法の点では助数詞の振る舞いが複雑多様であることを指摘した上で、その個別的な用法を用例に即して紹介している。

(見坊 1965) では世の中には生活の必要に応じていろいろな物の数え方があり、生活の新しい場面、新しい話題に即して新しい助数詞が生まれると述べられている。その一方で明治の始まる

前後までは助数詞として用いられてきたが、現在はその機能が失われている語の例が指摘されている。このような「助数詞の用いられ方の変化」に着目し、(荻野 1989) は大規模なアンケートを実施して、助数詞の実際の用いられ方について調査を行った。さまざまな年代、さまざまな職種の男女半数づつ計 425 人に対して行われたアンケート結果から年齢層別の助数詞の使われ方の差異を明らかにしている。さらに(谷原, 顔瑞珍, Debbie 1990) では(荻野 1989) の調査結果に基づき、5 つの助数詞の典型的な用法について考察している。

助数詞を数詞に接続する付属語ではなく数詞の語尾とする研究もある(森 1958)。しかし、本稿では二つの理由からこの立場は採用しない。一つは、助数詞に関する様々な現象を説明するのに、語尾という考え方は不適切であると考えからである。(見坊 1965) で指摘されているような多様性や可変性を説明するのに、数詞と助数詞は独立した語であると考えるのが自然である。もう一つは、LFG 規則を実装する日本語 LFG システムにおいて、形態素解析に茶釜(数詞とその語尾としての助数詞)を用いているためである。茶釜では数詞と助数詞を明確に分けている²。

本稿で扱う対象ではないが、日本語数量詞の大きな研究課題の一つに数量詞遊離構文の問題がある(岡部 2006)。遊離数量詞とは(a)「3 個のりんごを食べた」(b)「りんごを 3 個食べた」の(b)のような構文である。この構文をめぐって、(a)の連体数量詞文との置換の可否や、その決定要因、また(a)、(b)の表現の意味やニュアンスの違いについて様々な先行研究が分析や解釈を行っている。これらの課題を解決するためには、文法的な関係というよりはむしろ数量表現とそれが指す名詞や動詞などの意味関係に着目しなければならない。

2.2 助数詞の分類に関する研究

(Matsumoto 1993) はプロタイプ理論に基づいて、認知意味論的な立場から、助数詞の分類を試みている。(飯田 2004) は、助数詞ならびに助数詞と同じ働きをする名詞 600 語に対し、数えられる対象や同じ対象を数える助数詞群の共通点や差異について解説を行っている。(白井, 徳永 2007) は(飯田 2004)に記載されている助数詞 c とそれが数える名詞 n を書き起こしたデータから一つの c が二つ以上の n とペアになっているもの計 9,352 組を用いて、自動的に助数詞オントロジーを構築する手法について提案している。

2.3 助数詞を対象とした自然言語処理に関する研究

(Bond and Paik 2000) は翻訳などのアプリケーションに必要となる日本語文生成の際に、オントロジーを用いて名詞の意味的な情報を参照し助数詞を選択する手法について提案している。英語でも “ten feet long” の ‘feet’ や “two shelves higher” の ‘shelves’ のように日本語の助数詞に

² 但し、「一つ」、「二つ」のような数詞と助数詞「つ」、及び「一筆」、「四方八方」のような名詞として定着している数詞と助数詞については、まとめて一つの名詞としている。

似た数量表現が存在する。(Flickinger and Bond 2003) はこのような表現を対象にして HPSG システム上で解析を行う手法を提案している。

上記と同様に HPSG に関する研究ではあるが, 本稿に最も関連のある先行研究は (Bender and Siegel 2004) である。(Bender and Siegel 2004) は日本語 HPSG 文法において助数詞を解析し, 適切な解析結果, すなわち Minimal Recursion Semantics (以下, MRS) を出力する手法について提案している。本研究も基本的な目的は同様であるが, 次の点が異なる。まず, (Bender and Siegel 2004) では網羅的な助数詞の処理を提案していながら, MRS についての考察は主に対象となる名詞を数え上げるための助数詞に限定されていることである。本稿では, その他の助数詞についても考察を行う。さらに, 助数詞が省略されていてもその省略の有無に関係なく同一の f-structure を導出する文法や助数詞の性質を変化させる接頭辞や接尾語を考慮した文法について提案する。また, 二つの基準を設定して f-structure の妥当性を検討することも本稿の特徴である。さらに (Bender and Siegel 2004) には, 提案されている語彙規則 (type hierarchy) に助数詞をどのように割り当てるかの具体的な記述がない。本稿では, LFG の語彙規則 (lexical category) に語を割り当てる方法についても言及する。また, 本稿では, 今回提案した文法がどのくらい網羅的に実際のテキストを処理できるかを検証するための精度評価実験を行う。

3 助数詞のための日本語 LFG 規則

3.1 LFG 規則の表記法について

提案規則の詳細について述べる前に, この節では LFG 規則の表記方法について説明する。

LFG 規則は語彙規則と文法規則によって定義される。語彙規則では, f-structure の素性に対応する SUBJ(ect) や OBJ(ect) などの引数 (argument) や lexical category, 属性と属性値などを定義する。(1) に語彙規則の例を示す。1 行目は動詞「読む」に V という lexical category が割り当てられており, PRED(icate)「読む」は argument に SUBJ と OBJ という二つの引数を持つことが記述されている。2 行目には名詞「太郎」の lexical category が N であり PRED が「太郎」であることが記述されている。3 行目には名詞「本」の lexical category が N であり PRED が「本」であることが記述されている。4 行目には格助詞「が」の lexical category が PP であり, 主格を示す属性 NOM(inative) とその値「+」を持つことが記述されている。5 行目には格助詞「を」の lexical category が PP であり, 目的格を示す属性 ACC(usative) とその値「+」を持つことが記述されている。

- (1) 読む V PRED=‘読む <($\uparrow$ SUBJ)($\uparrow$ OBJ)>’
 太郎 N PRED=‘太郎’
 本 N PRED=‘本’
 が PP ($\uparrow$ NOM)=+
 を PP ($\uparrow$ ACC)=+
- (2) S $\rightarrow$ NP V
 $\{(\uparrow$ SUBJ)= $\downarrow$ ($\uparrow$ NOM)= $_c$ + | ($\uparrow$ OBJ)= $\downarrow$ ($\uparrow$ ACC)= $_c$ + $\}$ $\uparrow$ = $\downarrow$
- NP $\rightarrow$ N PP
 $\uparrow$ = $\downarrow$

文法規則は、句構造規則とそれに付与される機能的注釈で表現される。注釈において用いられる ' $\uparrow$ ' という記号は句構造の一つ上の節点に存在する f-structure を、' $\downarrow$ ' は現在の節点の f-structure を指す。(2) に (1) の語彙規則を含む文法規則を示す。1 行目の句構造規則では、0 個以上の NP と V が S になることを定義している。V に付与されている注釈は V の PRED が一つ上、すなわち文全体の PRED であることを定義している。NP に付与されている注釈は NOM 属性の値が '+' であれば NP の値が一つ上の SUBJ に、ACC 属性が '+' であれば OBJ になることを定義している。2 行目の句構造規則では、N と PP が NP になることを定義している。N に付与されている注釈は PRED が一つ上の節点である NP の PRED であることを定義している。(1) の語彙規則と (2) の文法規則を「太郎が本を読む」の解析に適用すると、図 1 に示すように句構造

図 1 LFG 規則の適用例

規則によって c-structure の木構造が形成されると同時に、機能的注釈によって属性と属性値で表現される f-structure が生成される。

3.2 助数詞のための語彙規則

表 1 に助数詞の語彙カテゴリの一覧を示す。本稿では助数詞が数詞に持たせる数の性質（基数／序数）という側面から「数量」, 「順番」という二つの語彙カテゴリを設定する。また, (国語学会 1980) では単位名を表す「メートル」, 「円」などを助数詞に含めないとする説があることを紹介しているが, 文法的には同じ働きをすることが指摘されている。本稿ではこの立場から, これらの語も助数詞とする。ただし, 上述の二つのカテゴリとは別に「単位」, 「通貨」という項目を設け, さらに数詞との位置関係から通貨を細分類した。それぞれの語彙カテゴリについて以下に述べる。

単位に関する語彙カテゴリは, 計測する対象の計測基準量を表す助数詞に割り当てられる。日本語 LFG システムにおいてこの語彙カテゴリには EDR 概念辞書 (日本電子化辞書研究所 1996) の接尾語のうち, 単位に分類されている語から「ヶ月」, 「週目」などの時間に関する語を除いた計 659 語を採用する。助数詞の中には単独で名詞として働く語も存在する。「身長をcmで, 体重はKgで入力して下さい」や「対ユーロでドルが下落している」など, 単位や通貨も数詞を伴わずに単独で名詞として用いられる場合がある。従って, このカテゴリの語には単に助数詞としての語彙ルールだけではなく, 名詞としてのカテゴリも割り当てる。さらに, 助数詞として働く場合でも名詞の属性である NTYPE を付与する。これは, 後述する英語の f-structure と平行にするためでもある。

通貨に関する語彙カテゴリは (Bender and Siegel 2004) と同様にその語彙的な特徴から二つに分ける。まず, 数詞の後に接続する後置詞, つまりいわゆる助数詞とほぼ同じふるまいをする語群である。この語彙カテゴリに属する語は EDR 概念辞書の「計量の単位」(概念 ID ‘30f93d’) の下位概念である「金銭の単位」(概念 ID ‘4446bd’) の直下にある語を採用した。ただし, これらの語のうち「両」や「文」など通貨の単位ではあるものの現在ほとんど使われておらず, 助

表 1 助数詞の語彙カテゴリ一覧

語彙カテゴリ	語数	属性	属性値	助数詞の例
単位	659	NTYPE	measure	kg, KG, km, キログラム
通貨 (後置)	135	NTYPE	currency	ドル, 円, ポンド, ペソ
通貨 (前置)	3	NTYPE	currency	\$, ¥, £
数量	76	CLASSIFIER-FORM	表層形	人, 個, 粒, 膳, 部
		CLASSIFIER-TYPE	cardinal	
順番	10	CLASSIFIER-FORM	表層形	位, 代目, 番手, 次
		CLASSIFIER-TYPE	ordinal	

数詞の他のカテゴリに分類した方が適切であると判断した語は除いた。一方、数詞の前に接続する前置詞のカテゴリにはコーパスを観察した結果得られた 3 語だけが割り当てられている。これらの語は助数詞というよりは純粋な名詞、もしくは記号としての性質が強い。しかし、通貨に関する数量表現の整合を考慮して、これらの語にも語彙カテゴリと文法ルールを割り当てた。通貨も単位と同様の理由から名詞のカテゴリも割り当て、NTYPE 属性も付与する。

数量に関する語彙カテゴリはもっとも典型的な助数詞と言える語に割り当てるカテゴリである。このカテゴリには、助数詞の表層を示す CLASSIFIER-FORM 属性と助数詞の性質を示す CLASSIFIER-TYPE 属性を付与する。このカテゴリに属する語は IPA 辞書 (IPAdic) に登録されている接尾語のうち、数えられるものを特徴づける語を人手で選別した。

順番に関する語彙カテゴリは物事の順序を示す助数詞である。このカテゴリにも CLASSIFIER-FORM と CLASSIFIER-TYPE 属性を与える。このカテゴリには、英語の序数を英訳したときに用いられる語である「等」、「位」、「号」、「級」、「流」及び IPA 辞書において品詞が助数詞として定義されている語のうち「目」、「番」、「次」を含む語で、単位を示す「貫目」、「次元」を除いた語を採用した。

3.3 助数詞のための文法規則

表 2 に助数詞の語彙カテゴリの一覧を示す。本節では数詞が助数詞もしくは名詞に接続して数量を表現している構文を解析するための文法について述べる。助数詞の語彙カテゴリは 3.2 節のものをを用いる。

3.3.1 単位や通貨に関する文法規則

単位や通貨に関する文法規則を (3) に示す。単位や通貨を表す助数詞が数詞と接続する場合には、助数詞を PRED とし、数詞がそれを修飾する。これは英語の “3 kg” や “10 dollars” の f-structure との整合性を図ると同時に、3.2 節でも述べたようにこれらの助数詞が接尾辞という

表 2 助数詞を解析するための文法規則一覧

解析対象	構文例	解析に必要な文法規則
単位／通貨	「300 円」, 「300 kg」	(3)
通貨	「¥300」	(4)
数量	「りんご 3 個」, 「3 人の学生」	(5)
	「3 個」, 「30 人の協力」	(6) (7)
	「3 家族」	(8)
順番	「3 番目」	(9)
	「3 人目」	(10)
	「第 2 の容疑者」	(11)

よりもむしろ独立した名詞としての性質が強いと考えるからである。

単位・通貨は修飾先の名詞の属性について述べるために用いられる。例えば「60 デニールのストッキング」であれば「60 デニール」はストッキングの数量ではなく、「繊維密度」を表現しているのであり、「300 円のストッキング」の「300 円」はストッキングの「価格」を表現している。つまり「60 デニール（という繊維密度）のストッキング」、「300 円（という価格）のストッキング」と解釈できる。従って、「60」や「300」などの数詞は直接「ストッキング」を修飾しているのではなく、「デニール」や「円」を修飾しているとして解析するのが妥当である。

さらに、通貨を示す接頭語が数詞に接続する表現を解析する文法を (4) に示す。これも上記と同じ理由から接頭語を PRED とし、数詞がそれを修飾する。

- $$(3) \quad \text{Nadj} \longrightarrow (\text{NUMBER}) \quad \{ \text{CL_unit} \mid \text{CL_currency} \}$$
- $$(\uparrow \text{SPEC NUMBER}) = \downarrow \quad (\uparrow \text{PRED}) = \downarrow$$

Nadj: 体言 NUMBER: 数詞 CL_unit: 単位に関する助数詞 CL_currency: 通貨に関する助数詞 (): 省略可能 |: OR PRED: 主辞 SPEC: 指定部

- (4) Nadj $\longrightarrow$ CL_Pref.currency NUMBER
 ($\uparrow$ PRED)= $\downarrow$ ($\uparrow$ SPEC NUMBER)= $\downarrow$

CL_Pref_currency: 通貨に関する接頭語

3.3.2 数量に関する文法規則

数量を表す助数詞は、数詞の中でも「もの」や「こと」の数量を表現するのに用いられる典型的なものである。「3個のりんご」や「3箱のりんご」の「3個」や「3箱」はいずれもりんごの数や量を表している。単位・通貨とは異なり、数詞「3」は「りんご」を直接修飾していると考えられる。つまり、助数詞は「数え方」を表現するために補助的に用いられる接尾語として扱う。この場合、「個」と「箱」の数え方の違いは f-structure の構造そのものには反映されないが、助数詞の表層を属性値として表現することで、オントロジなどの知識源を利用した意味解析や文脈解析の処理を利用してより詳細な数え方の違いを扱うことを想定している。

数量に関する文法規則は、数詞と助数詞だけではなく、それによって数えられる名詞についても記述する必要がある。例えば「りんご3個」と「3個のりんご」はそれぞれ「3個」が「りんご」の数を表しているため、これらの解析結果は同じであることが望ましい。

(5)に「3個のりんご」と「りんご3個」を解析するための文法規則を示す。この規則ではPREDはりんごであり、その数が3であると解析する。これは英語の“3 apples”と同じf-structureを持たせることを意図した分析である。

- (5)
$$\begin{array}{lcl} \text{Nadj} \longrightarrow & \{ \text{NPnum} & \text{Nadj} \\ & (\uparrow \text{SPEC NUMBER})=\downarrow & (\uparrow \text{PRED})=\downarrow \\ & | \text{Nadj} & \text{Nnumeric} \} \\ & (\uparrow \text{PRED})=\downarrow & (\uparrow \text{SPEC NUMBER})=\downarrow \end{array}$$
- $$\text{NPnum} \longrightarrow \text{Nnumeric} \quad \text{PPadnominal}$$
- $$(\uparrow \text{PRED})=\downarrow$$
- $$\text{Nnumeric} \longrightarrow \text{NUMBER} \quad \text{CL_cardinal}$$
- $$(\uparrow \text{PRED})=\downarrow \quad (\uparrow \text{NUMBER-TYPE})=\text{cardinal}$$

CL_cardinal: 数量に関する助数詞 PPadnominal: 連体化助詞

また「3 個の値段」、「3 個を食べた」など、数える対象である名詞が省略され、数詞と助数詞が単独で現れる場合もある。このような現象を解析するための文法規則を (6) に示す。この文法規則では PRED には代名詞を示す記号 ‘PRO’ を挿入し、その属性に表層が無いことを示す属性値 null を与える。この操作によって、名詞の省略の有無によらず同じ構造の f-structure を出力することができる。

上述 (5) と下記の (6) の二つの規則は数詞と助数詞に同時に適用される。体言には (7) に示すように連体化助詞による連体修飾のための規則も定義されていることから、数量に関する表現が連体化助詞によって連体修飾を行う場合には、常に二つの f-structure が出力される。

- (6)
$$\begin{array}{lcl} \text{Nadj} \longrightarrow & \text{Nnumeric} & \text{e} \\ & (\uparrow \text{SPEC NUMBER})=\downarrow & (\uparrow \text{PRED})=\text{'pro'} \\ & & (\uparrow \text{PRON-TYPE})=\text{null} \end{array}$$

e: 省略記号

- (7)
$$\begin{array}{lcl} \text{Nadj} \longrightarrow & \text{NPadnominal} & \text{Nadj} \\ & (\uparrow \text{ADJUNCT})\ni\downarrow & (\uparrow \text{PRED})=\downarrow \end{array}$$
- $$\text{NPadnominal} \longrightarrow \text{Nadj} \quad \text{PPadnominal}$$
- $$(\uparrow \text{PRED})=\downarrow$$

ADJUNCT: 修飾成分

また、上述とは逆に、「3 自衛隊」や「7 競技」のように助数詞を伴わずに直接数詞が名詞に接続してその数を表すことがある。このような現象を解析するための文法規則を (8) に示す。この文法規則では助数詞の表層を表す CLASSIFIER-FORM 属性に表層が無いことを示す属性値 ‘null’ を与えることにより、助数詞の有無によらず同じ構造の f-structure を出力することができる。

- (8) Nadj $\rightarrow$ NUMBER e N
 ($\uparrow$ SPEC NUMBER)= $\downarrow$ ($\uparrow$ CLASSIFIER-FORM)=null ($\uparrow$ PRED)= $\downarrow$

N: 名詞

3.3.3 順番に関する文法規則

数詞が順番を示す表現は主に三つが想定される。まず一つは 3.2 節で挙げた順番に関する助数詞が接続した場合である。このような構文に対しては (5) の文法がほぼそのまま適用できる。もう一つは接尾辞や接頭辞の接続によって数詞や助数詞の性質が変化する場合である。これは、「3 人目」、「5 回目」のように数量を表す助数詞に接尾辞「目」が接続する場合と「第 3 章」、「第 1 回」など数詞の前に接頭辞「第」が接続する場合がある。さらに、「第 3 回目」、「第 5 次」など、接頭辞は助数詞が順番を示していても任意に接続する場合がある。

これらの現象を解析するための文法規則を (9) に示す。助数詞の属性値が基数を示す cardinal であっても、数詞の示す数の性質は基数から序数に変化する。従って、語彙ルールには接尾辞によって数詞の属性 NUMBER-TYPE が序数を示す値 ordinal になる可能性について記述しておく必要がある。(10) に数量を示す助数詞「回」の語彙規則の記述例を示す。

- (9)
Nnumeric $\rightarrow$	CL_Pre_ord	NUMBER	CL_cardinal
	{($\uparrow$ NUMBER-TYPE)= ordinal ($\uparrow$ PRED)= $\downarrow$		
	($\uparrow$ NUMBER-TYPE)= _c ordinal}		CL_Post_ord
			($\uparrow$ NUMBER-TYPE)
= ordinal
- CL_Pre_ord: 第 CL_Post_ord: 目
- (10) 回 CL_cardinal {($\uparrow$ NUMBER-TYPE) = cardinal
 | ($\uparrow$ NUMBER-TYPE)=_c ordinal}
- _c: 制約条件の相等

さらに接頭辞「第」は「第 3 の男」、「第 6 のコース」のように助数詞を伴わない単独の数詞に接続して数の性質を序数にする場合がある。このような現象を解析するための文法規則を (11) に示す。

- (11) Nadj $\rightarrow$ CL_Pre_ord NUMBER
 ($\uparrow$ NUMBER-TYPE)=ordinal ($\uparrow$ PRED)= $\downarrow$

3.3.4 機能的関係と照応的關係について

(Bender and Siegel 2004) は、可算名詞を修飾する助数詞の処理において、「猫を 2 匹飼う」のように、格助詞「を」に接続する NP の後に単独で現れる数詞と助数詞を NP の修飾要素とし

て扱う文法規則を提案している。しかし、本稿の文法規則はこのように数詞と助数詞が単独で現れる場合には助数詞の修飾先として代名詞を示す記号‘PRO’を挿入する。これは、「2匹」と「猫」の関係が統語機能的な修飾関係にあるのではなく、照応関係にあると考えるからである。(Bresnan 2001)では、f-structureでは統語機能的関係を表現すべきであり、照応的な関係はその対象としないと述べているが、本稿もこの立場から文法規則を設計した。

日本語構文では数量表現が指すNPが必ず数量表現の直前に現れるとは限らない。例えば「猫をマンションで2匹飼っていた」のように、「2匹」の係り先が構文中の任意の場所に現れる可能性がある。(Bender and Siegel 2004)の文法規則では修飾先のNPが直前に現れない場合は数量詞の修飾を特定しない。つまり本稿と同様に照応的な関係として解析する。しかしその結果、「猫をマンションで2匹飼っていた」と「マンションで猫を2匹飼っていた」の解析結果が異なることになる。本稿では、こういった単なる語順の違いにf-structureが影響されないことを配慮している。

4 f-structureの妥当性の検討

4.1 日本語LFGシステムの構成

LFG規則の妥当性の検証と解析精度の評価実験を行う処理は(増市, 大熊 2003)に改良を加えて拡張した日本語LFG解析システムを用いる。図2に、本稿で用いる日本語LFGシステムの構成を示す。まず日本語入力文が日本語taggerに渡される。日本語taggerではChaSen (cha)による形態素解析が行われる。その結果に表層文字列や品詞情報を利用した形態素区切りの修正ルールを適用し、修正された形態素列に品詞情報を利用したアノテーションルールによってtagを付与する。

tagが付与された形態素列をXerox Linguistic Environment (XLE)に入力する。XLEはLFG理論の仕様をほぼ完全に実装した処理系である(Maxwell and Kaplan 1993)。日本語システムは大きく分けて3種類のデータで構成されている。一つはメモリ容量や探索深さ、LFG規則の優先順位などの設定を定義するデータ群である。二つ目はLFGの文法理論に沿って実装された文法規則群である。3.3章で提案した文法規則群は名詞句に関する文法規則に追加されている。三つ目は語彙規則群である。3.2章で提案した助数詞に関する語彙規則はここに追加されている。

4.2 他表現のf-structureとの比較

本節では3章で提案したLFG規則が導出するf-structureが1.2節で設定した基準1を満たしているかを検証するために、同じ統語意味構造を持ちながらも語順の変化や省略によって表層形が異なった構文のf-structureを比較する。

一般に通貨や単位は数詞を伴わないで単独の名詞として用いられる例も多い。そこで通貨の

図2 日本語 LFG システムの構成

図3 「100ドルを両替した。」と「ドルを両替した。」のf-structure

助数詞が数詞を伴う構文「100ドルを両替した。」に対して文法規則(5)を適応して導出したf-structureと数詞を伴わない構文「ドルを両替した。」のf-structureを図3に示す。両者を比較すると同じ構造を持っていることが分かる。

次に数量に関する表現のf-structureについて検討する。数量を表現する文法規則は3.3節の文法規則(5)と文法規則(6)があり、これらは同時に適用されるので必ず2通りのf-structureが出力される。「3箱の煙草」について図4に文法規則(5)によって導出されたf-structureを図5に文法規則(6)によって導出されたf-structureを示す。図4のf-structureで「煙草」のSPEC

図 4 文法規則 (5) によって導出された「3 箱の煙草」の f-structure

図 5 文法規則 (6) によって導出された「3 箱の煙草」の f-structure

図 6 文法規則 (8) によって導出された「3 銘柄」の f-structure

の NUMBER 属性中の PRED が「3」になっているのは、煙草の数量が3であることを表現している。従って、正しい f-structure は文法規則 (5) によって導出された図 4 の f-structure である。

さらに、助数詞を伴わず直接数詞が名詞を修飾してその数を表す場合も、文法規則 (8) によって図 4 と同じ構造の f-structure で表現される。図 6 に「3 銘柄」の f-structure を示す。

しかし、数詞の数える対象が係り先の名詞ではない場合がある。例えば、「3 箱の値段」という表現において数詞「3」の数える対象は「値段」ではない。この名詞句の解釈は「3 箱の何か (例えば煙草) の値段」とするのが妥当である。図 7 と図 8 に「3 箱の値段」の f-structure を示す。文法規則 (6) によって導出された図 8 の f-structure では表層が省略されていることを現す

図 7 文法規則 (5) によって導出された「3 箱の値段」の f-structure

図 8 文法規則 (6) によって導出された「3 箱の値段」の f-structure

PRON-TYPE の属性値 null を持つ代名詞 'pro' の数量として表現されている。つまり、図 8 の f-structure が「3 箱の値段」の正しい統語意味構造を表している。また、数量表現が係り先の名詞を伴わずに単独で現れる場合も (6) によって適切な f-structure を出力する。例として「彼が選んだ 3 箱」の f-structure を図 9 に示す。

次に、順序を示す表現について検討する。それぞれの表現の差異は、属性や属性値によって表現されるが、基本的な構造はすべて等しいことが分かる。

図 10 に文法規則 (9) によって導出された「3 番目の男」の f-structure を示す。「番目」は順番を示す助数詞の語彙カテゴリが割り当てられているため、CLASSIFIER-TYPE 属性と NUMBER-TYPE 属性の値は共に序数を示す ord(inal) になっている。

図 11 に文法規則 (10) によって導出された「3 人目の男」の f-structure を示す。「人」は数量を示す助数詞の語彙カテゴリが割り当てられているため、CLASSIFIER-TYPE 属性の値には基数を示す card(inal) が入る。しかし、接尾辞「目」が順番を表現しているため NUMBER-TYPE 属性の値が序数を示す ord(inal) になっている。

図 12 に文法規則 (11) によって導出された「第 3 の男」の f-structure を示す。この名詞句には助数詞が存在せず、助数詞の省略も想定され難いため CLASSIFIER-TYPE 属性を持たない。し

図 9 文法規則 (6) によって導出された「彼が選んだ3箱」の f-structure

図 10 文法規則 (9) によって導出された「3番目の男」の f-structure

図 11 文法規則 (10) によって導出された「3人目の男」の f-structure

かし、接頭辞「第」が順番を表していることを NUMBER-TYPE 属性の値が ord(inal) になっていることで示している。

図 12 文法規則 (11) によって導出された「第3の男」の f-structure

図 13 “I exchanged 100 dollars” の f-structure

4.3 他言語の f-structure との比較

本節では 1.2 節で設定した基準 2 について検証するために、日本語の表現に対応する他言語の f-structure と日本語の f-structure の比較を行う。

まず通貨・数量に関する表現の f-structure について検討する。図 13 に「100 ドルを両替した。」の英訳 “I exchanged 100 dollars.” を入力として (Riezler et al. 2002) の英語 LFG システムで出力された f-structure を示す。4.2 節の図 3 の日本語の f-structure と比較すると同じ構造を持っていることが分かる。

次に数量に関する表現の f-structure について中国語の f-structure との比較を行う。数量に関する表現は助数詞を伴う場合と助数詞が省略される場合がある。助数詞を伴う表現については (Fang and King 2007) の中国語 LFG システムで出力された f-structure と、助数詞を伴わない表現については英語の f-structure と比較する。

図 14 に文法規則 (5) が導出した「三本の酒」の f-structure を、図 15 に「三瓶酒」に対応する中国語の f-structure を示す。図 15 の f-structure では助数詞のための属性である CLASSIFIER

図 14 文法規則 (5) によって導出された「三本の酒」の f-structure

図 15 「三瓶酒」の f-structure

図 16 文法規則 (6) によって導出された「彼女は三本を飲んだ。」の f-structure

を PRED にしている。これは、中国語の助数詞が連体修飾を受けることができるほど名詞としての性質が強いためである。このような違いはあるものの、この表現に対する日本語と中国語の f-structure の基本的構造はほぼ同じである。

次に、助数詞が数える対象である名詞が存在しない数量表現について「彼女が三本を飲んだ」と「她喝了三瓶」の解析結果を比較する。図 16 に日本語の f-structure を、図 17 に中国語の f-structure を示す。両者とも数える対象に、代名詞を示す 'pro' を代入している。これは、数える対象の名詞が存在する場合の f-structure との整合性を図るための手当てであるが、それと

```

"她/r 喝/v 了/u 三/m 瓶/q . /w "
[PRED '喝<[1:她], [654:pro]>']
SUBJ [PRED '她']
      1[ANIM '她', NUM sg, PERS 3]
      [PRED 'pro']
OBJ [SPEC [CLASSIFIER 107[PRED '瓶']
           NUMBER 82[PRED '三',
                     NTTYPE [NSYN number]]]
      654NUM pl
TNS-ASP [ASPECT N-ASPECT perfective]
28[CLAUSE-TYPE decl, VTYPE main]

```

図 17 「她喝了三瓶」の f-structure

図 18 「3 家族」と「3 families」の f-structure

図 19 “third man” の f-structure

時に中国語と日本語の間の f-structure の整合性を保持することも可能にしている。

助数詞が省略される表現については英語の f-structure と比較を行う。図 18 に「3 家族」と「3 families」の f-structure を示す。日本語の f-structure では助数詞が存在しないことを明示的に表現するために CLASSIFIER-FORM 属性が null になっているが、英語では基本的に助数詞を用いないので、それに対応する属性は存在しない。しかし、基本的な構造は等しいことが分かる。

次に順序を示す表現について検討する。図 19 に 4.2 節の図 10～12 に対応する英語 “third man” の f-structure を示す。特に図 12 の「第 3 の男」の f-structure とは同じ構造を持っている。

5 解析結果の精度評価

5.1 評価方法

解析結果の評価は f-structure を (King, Crouch, Riezler, Dalrymple, and Kaplan 2003) に準拠した形式 (triples) に変換して行う。この変換の際に文法的な情報ではなく実装の都合上付与されている属性は削除する。図 20 に「3 個のりんご」の f-structure と triples の対応関係を示す。

triples 形式は“属性 C (ノード A , ノード B)”もしくは“属性 C (ノード A , 属性値 B)”という書式で表現される。いずれも「 A の属性 C は B である」と解釈する。属性の要素はノードか値であるが、ノードは必ずノード ID を持つため、同じ語が同一文中に現れても区別が可能である。図 20 中の triples 形式のデータは「りんご」の NUMBER 属性が 3 であること、「3」の CLASSIFIER-FORM 属性の値は「個」であること、「3」の CLASSIFIER-TYPE 属性の値が cardinal であること、「3」の NUMBER-TYPE 属性の値が cardinal であることを表現している。

解析対象は EDR コーパスから数詞を含む文を無作為に選び、単位・通貨に関する表現、数量に関する表現、そして順序に関する表現を手で 200 表現ずつ抽出した。

本稿で提案する規則の効果を確認するために、精度の比較を (増市, 大熊 2003) で採用されている LFG 規則の解析結果と行う。今回の提案規則による効果を正確に計測するために、旧規則の解析結果に単なる仕様変更によって正解でない NUMBER-TYPE などの属性が含まれる場合はこの属性を削除した。評価実験における triples に含まれる属性の一覧を表 3 に示す。

今回の実験では、多義性の解消を行っていないため、複数の解析結果が出力される可能性があるが、精度の測定はそれらの解析結果すべてに対して行った。例えば、4.2 節で「3 箱の煙草」に対して得られた図 4 と図 5 の二つの解析結果のうち、図 4 が正解であるならば、この解析結

図 20 f-structure と triples の対応関係

表 3 triples に含まれる属性一覧

属性	説明	属性	説明
adjunct	修飾成分	det	限定詞
atype	形容詞の種類	fragment	部分解析
classifier_form	助数詞の表層	mod	複合名詞の主辞以外の要素
classifier_type	助数詞の種類	number	数
clause_type	節の種類	obj	目的語
conj	並置	pron_type	代名詞の種類
coord_form	接続詞の表層	subj	主語
coord_type	接続詞の種類	tense	時制

表 4 「3 箱の煙草」の解析結果

Dependency	Precision	Recall	FScore
adjunct	0/1 = 0	0/0 = 0	0
classifier_form	2/2 = 100	2/2 = 100	100
classifier_type	2/2 = 100	2/2 = 100	100
number	2/2 = 100	2/2 = 100	100
pron_type	0/1 = 0	0/0 = 0	0

果の精度は下記の表 4 のように求められる。

今回提案した LFG 規則のみを評価するため、解析する範囲を数詞、助数詞及びそれらが連体連体修飾を行っている場合はその修飾先の名詞句、連体修飾を受けている場合にはその修飾成分のみに限定した。また「およそ 30 個」の「およそ」や「5%以下」の「以下」など直接接続して助数詞や数詞の修飾を行う付属語的な名詞も対象とした。下記に解析対象の例を示す。

- 単位・通貨に関する表現
 - － 400 万円のミンクのコート、250 万円の腕時計も頭にちらついた。
 - － IBM およびコンパチはパソコン全体の 80 %以上を占めた。
- 数量に関する表現
 - － 3 国合わせての総人口が約 750 万人だから 5 人に 1 人が参加したことになる。
- 順序に関する表現
 - － 2 番目の LANは事務職員のための LAN だ。
 - － AST の第 1 のライバルである強化ボードの開発者クォードラムでは、IBM が新しいラインに残したメモリー拡張のギャップに焦点を合わせている。

5.2 結果と考察

5.2.1 単位・通貨に関する表現

表5に単位・通貨に関する表現の解析結果を、表6と表7に属性別の結果を示す。旧規則では単位・通貨と他の助数詞との区別をしていなかったため、CLASSIFIER-FORM属性とCLASSIFIER-TYPE属性が解析結果に含まれる。そこで、今回の比較は、この二つの属性を予め削除して行った。今回提案したLFG規則により全体的なF値はおよそ25%向上した。表6のfragmentは、部分解析結果の属性であり、旧規則では適用される規則が存在せず部分解析しかできなかった表現があることを示している。表7にこのfragment属性が存在しないことは提案規則ではすべ

表 5 単位・通貨に関する表現の解析結果

LFG 規則	precision	recall	f-score
旧規則	52.56%	51.59%	52.10%
提案規則	76.68%	79.14%	77.90%

表 6 単位・通貨に関する表現の属性別の解析結果（旧規則）

Dependency	Precision	Recall	FScore
adjunct	57/104 = 55	57/142 = 40	46
atype	2/2 = 100	2/2 = 100	100
clause_type	2/2 = 100	2/2 = 100	100
conj	0/0 = 0	0/12 = 0	0
fragment	0/4 = 0	0/0 = 0	0
mod	70/108 = 65	70/78 = 90	75
number	58/102 = 57	58/148 = 39	46
pron_type	2/31 = 6	2/4 = 50	11
subj	2/2 = 100	2/2 = 100	100
tense	2/2 = 100	2/2 = 100	100

表 7 単位・通貨に関する表現の属性別の解析結果（提案規則）

Dependency	Precision	Recall	FScore
adjunct	144/203 = 71	144/213 = 68	69
atype	2/6 = 33	2/2 = 100	50
clause_type	2/6 = 33	2/2 = 100	50
conj	0/0 = 0	0/12 = 0	0
mod	73/103 = 71	73/94 = 78	74
number	216/232 = 93	216/241 = 90	91
pron_type	3/7 = 43	3/3 = 100	60
subj	2/6 = 33	2/2 = 100	50
tense	2/6 = 33	2/2 = 100	50

て全体的な解析に成功していることを示している。また, EDR 概念辞書を用いて語彙エンタリを追加したことにより, 修飾関係を示す属性 adjunct の F 値が向上している。表 7 で subj, tense, atype, clause_type の precision の値が下がっているのはすべての単位・通貨に名詞の lexical category を割り当てた結果, 文法規則 (8) が誤って適用されてしまう例があったためである。今後は一般的に名詞としても用いられる単位名とほとんど名詞としては現れない単位の分別が必要である。

併置関係を示す属性 conj の F 値が両規則ともに 0 なのは, 下記の例にあるような数詞と記号による表現を提案規則でカバーできなかったためである。特に単位に関する表現は記号を含む表現が多く見られ, 今後はこれらの記号と数詞に関する規則の精緻化が課題である。

- 640 × 400 ドットの解像度でグラフィックを表示できるという。
- 旧システムの 30 % ~ 50 %をリプレイスする大規模な計画で, 2 年がかりで構築する。

5.2.2 数量に関する表現

表 8 に数量に関する表現の解析結果を, 表 9 と表 10 に属性別の結果を示す。全体的な F 値は

表 8 数量に関する表現の解析結果

LFG 規則	precision	recall	f-score
旧規則	85.04%	79.03%	81.92%
提案規則	87.74%	86.91%	87.32%

表 9 数量に関する表現の属性別の結果 (旧規則)

Dependency	Precision	Recall	FScore
adjunct	63/99 = 64	63/105 = 60	62
atype	13/13 = 100	13/13 = 100	100
classifier_form	176/177 = 99	176/192 = 92	95
classifier_type	137/137 = 100	137/143 = 96	98
clause_type	13/13 = 100	13/13 = 100	100
conj	3/6 = 50	3/6 = 50	50
coord_form	2/3 = 67	2/2 = 100	80
coord_level	2/3 = 67	2/2 = 100	80
det	1/1 = 100	1/1 = 100	100
mod	40/64 = 62	40/41 = 98	76
number	110/130 = 85	110/192 = 57	68
obj	1/8 = 12	1/2 = 50	20
pron_type	107/122 = 88	107/125 = 86	87
subj	1/13 = 8	1/13 = 8	8
tense	13/13 = 100	13/13 = 100	100

表 10 数量に関する表現の属性別の結果 (提案規則)

Dependency	Precision	Recall	FScore
adjunct	76/117 = 65	76/129 = 59	62
atype	7/7 = 100	7/7 = 100	100
classifier_form	222/224 = 99	222/226 = 98	99
classifier_type	157/157 = 100	157/159 = 99	99
clause_type	7/7 = 100	7/7 = 100	100
conj	2/4 = 50	2/4 = 50	50
coord_form	1/2 = 50	1/1 = 100	67
coord_level	1/2 = 50	1/1 = 100	67
det	4/5 = 80	4/5 = 80	80
mod	34/45 = 76	34/36 = 94	84
number	182/224 = 81	182/226 = 81	81
obj	1/4 = 25	1/2 = 50	33
pron_type	122/134 = 91	122/138 = 88	90
subj	7/7 = 100	7/7 = 100	100
tense	7/7 = 100	7/7 = 100	100

およそ 5 % 向上している。両者とも連体修飾を示す adjunct 属性の値が低いのは、文法規則 (5) と文法規則 (6) によって生じた多義性によるものである。この種の多義性を解消するためには (白井, 徳永 2007) や (Bond and Paik 2000) で提案されているオントロジーのような知識資源の利用が有効であると思われる。その一方で、下記の例の「2 人」と「仲間」のように、助数詞と数えられる対象の関係だけでは正しい f-structure を決定することが困難である場合もあり、文脈解析のようなより深い知識処理が必要である。

- 3 年前には、横浜防衛施設局に仲間 2 人が火炎瓶を投げて逮捕された。
- 2 人の仲間だったシンイチは、いつか姿を消した。

さらに、連体修飾成分と数詞、助数詞の関係を表現するのが難しい場合があった。下記の文の「長男と 2 人」は「長男と (自分の) 2 人」を指しており、「長男」と「2 人」の並置と考えるよりは「長男と」が「2 人」を修飾すると解釈するのが適切であるが、本稿で提案した LFG 規則ではこのような表現は扱えない。

- 10 年前、姉妹を頼って、長男と 2 人で引き揚げてきた。

また、本稿で提案した規則では評価用コーパス中下記の表現が解析できなかった。「100 万票差」の「差」、あるいは「合計」、「積」なども含めて数量操作に関わる名詞については今後 f-structure の表現方法も含めて扱い方を検討する必要がある。

- 71 年の大統領選挙で朴大統領に約 100 万票差で敗れる。
- 11 の市民団体が実行委員会を結成し、それぞれの企画を原爆忌の 9 日まで繰り広げる。

- これで輸入解禁州は全部で 38 になる。

上記の「11 の市民団体」の「11」や「全部で 38 になる」の「38」のように、助数詞や接頭辞、接尾辞など数詞の性質を表現する機能語が接続せずに名詞を修飾している場合、その数詞が何を表現しているのか判断するのが難しい。評価コーパスには数量を表す例しか出現しなかったが、下記の例文の「二百万のお宝」の「二百万」は「お宝」である「トルエン」の価格を表現しており、「一の富」の「一」は宝くじの一等を示している。

- 千本の物が一。二百万のお宝が。そのとき、トルエンが爆発した。[大沢在昌「ちきこん」『らんぼう』新潮文庫 (1998)]
- ふつうこれは百回ついたものだそうで、百回目を突きどめといって、これがつまり一の富、千両もらえるわけである。[横溝正史「鶴の千番」『くらやみ婿』春陽文庫 (1984)]

5.2.3 順序に関する表現

表 11 に順序に関する表現の解析結果を、表 12 と表 13 に属性別の結果を示す。旧文法では CLASSIFIER-TYPE 属性の序数を表現する値 ordinal を定義していなかったため、解析結果を公平に比較するためにこの属性を削除した。それでも、全体的な F 値は 21% 以上向上した。

ただし、本稿で提案した規則では扱えない表現も見られた。順序に関する表現は、通貨・単位や数量に関する表現よりも、助詞を介さずに直接他の名詞と連結して、複合名詞を形成する場合が多く、下記のように係り先がうまく決定できない例が散見された。本稿で用いた日本語 LFG システムでは、処理時間を高速化するために (特許 2007) の手法を用いて、名詞が二つ以上

表 11 順序に関する表現の解析結果

LFG 規則	precision	recall	f-score
旧規則	56.51%	63.09%	59.62%
提案規則	81.90%	80.63%	81.26%

表 12 順序に関する表現の属性別の結果 (旧規則)

Dependency	Precision	Recall	FScore
adjunct	52/205 = 25	52/101 = 51	34
classifier_form	133/133 = 100	133/180 = 74	85
conj	20/35 = 57	20/33 = 61	59
coord_form	12/14 = 86	12/12 = 100	92
coord_level	12/14 = 86	12/12 = 100	92
mod	125/246 = 51	125/132 = 95	66
number	70/103 = 68	70/220 = 32	43
obj	0/1 = 0	0/1 = 0	0
pron_type	58/96 = 60	58/79 = 73	66

表 13 順序に関する表現の属性別の結果 (提案規則)

Dependency	Precision	Recall	FScore
adjunct	49/69 = 71	49/65 = 75	73
classifier_form	149/158 = 94	149/187 = 80	86
conj	6/7 = 86	6/7 = 86	86
coord_form	3/3 = 100	3/3 = 100	100
coord_level	3/3 = 100	3/3 = 100	100
mod	87/111 = 78	87/96 = 91	84
number	159/210 = 76	159/219 = 73	74
obj	1/1 = 100	1/1 = 100	100
pron_type	59/62 = 95	59/65 = 91	93

連続するときにはすべて最右の名詞を主辞とする複合名詞としてまとめる処理を実施する。その結果、下記の (a) では「第 2 次」が複合名詞中で最も右にある「撤兵」を修飾するため正しい f-structure を得ることができる。しかし、(b) では「2 次」が複合名詞中の「下請け」を修飾しているので本稿のシステムでは誤った f-structure を導出する。

(a) このサマー地区はベトナム軍が第 2 次カンボジア撤兵の際に通過した地だ。

(b) 2 次下請け工場になると、忙しさだけは 1 次下請けに負けない。

また、下記の「版」、「条」、「面」のように助数詞が数量を現すものであっても実際には順序に関する表現である例があった。これは (荻野 1993) で示唆されているように、助数詞には基数と序数の両方を表すものがあるためであると思われる。今後は、こういった助数詞の特定も必要である。

- この批評は大阪本社発行の 13 版をもとにしています。
- 航空法施行規則 145 条に規定された各種の航空計器および航法装置。
- これは先日、離婚請求権が有責者にも認められたという判決が、新聞の 1 面に出た翌日の会話である。

6 おわりに

6.1 まとめ

本稿では数詞と助数詞によって表現される構文を日本語 LFG システムで解析するための語彙規則と文法規則を提案した。さらに提案した規則によって導出される f-structure を他の表現や他の言語の f-structure と比較して解析結果の妥当性を検討した。また、精度評価実験を実施して提案規則の解析能力について、従来の LFG 規則との比較を行った。その結果、通貨・単位に関する表現では 25%、数量に関する表現では 5%、順序に関する表現では 21%の F 値の向上が

確認できた。

6.2 今後の課題

3 章では、助数詞を通貨・単位、数量、順序という三つの語彙カテゴリに分類し、それぞれに対して文法規則を提案した。しかし、助数詞の中にはこの三つのカテゴリのどれか一つに明確に分類できないものも存在する。例えば、本稿で「組」は数を表す助数詞として分類されるが、「ダース」は単位を表す助数詞として分類されている。このように、単位が数や量を表す場合には今回の分類を明確に適用することが難しい。また、5 章の実験結果にも見られたように、「版」や「条」など数量と順序の両方を表す助数詞も存在する。このような問題を解決するためには、特殊な性質を持つ助数詞を特定し、本稿で提案した分類の細分化や複数の語彙カテゴリの割り当てなどを検討する必要がある。

4.3 節では本稿で提案した LFG 規則が導出する f-structure の妥当性を検証するために、英語と中国語の f-structure との比較を行った。これは、両者とも実用的な LFG システムが既に構築されており、日本語文の訳文に対して適切な f-structure を得ることが可能だったためである。しかし、他の言語の LFG システムの研究開発が進めば、他の言語の f-structure とも比較したい。

5 章の評価実験では、解析精度の比較対象として旧 LFG 規則の解析結果を用いたが、より正確に解析能力を検証するためには、他のシステムの出力結果と比較する必要がある。しかし、1.1 節でも述べたとおり、現在公開されている構文解析システムでは本稿が対象にしている助数詞と数詞、名詞の関係を表現しないため、f-structure と直接比較することができない。今後、翻訳や質問応答などのシステムの出力を利用して間接的に比較を行うなど、評価の工夫が必要である。本稿の評価実験対象は助数詞に関する表現に限定した。今回の改良によって文全体のカバー率には特に変化が見られなかったが、個々の解析結果が文全体の解析精度にどのような影響を及ぼしているかはより詳細に観察する必要がある。

今後は、今回解析できなかった構文を扱うために LFG 規則の改良と拡張に取り組む。また、(Umemoto 2005) や (Crouch 2005) など f-structure を入力とする意味解析手段、つまり LFG 解析の次の段階の処理を用いて、提案した規則によって生じた f-structure の曖昧性解消を実施したい。

謝 辞

ParGram のメンバー、特に助数詞の解析に関する議論を通じて有益なコメントをいただいたり、解析結果を提供していただいた Microsoft 社の Tracy Holloway King 氏、Martin Forst 氏、Palo Alto Research Center Inc. の Ji Fang 氏に感謝いたします。また、日本語 LFG について有益なコメントをいただいた早稲田大学の原田康也教授、中野美知子教授に感謝いたします。XLE

の開発者であり、日本語システム構築時に実装に関する貴重な助言を頂いた Palo Alto Research Center Inc. の John Maxwell 氏、Microsoft 社の Ronald Kaplan 氏に感謝いたします。

参考文献

- “ChaSen.” <http://sourceforge.jp/projects/chasen-legacy/>.
- “IPAdic.” <http://sourceforge.jp/projects/ipadic/>.
- Bender, E. M. and Siegel, M. (2004). “Implementing the Syntax of Japanese Numeral Classifiers.” In *IJCNLP-04*.
- Bond, F. and Paik, K. (2000). “Reusing an ontology to generate numerical classifiers.” In *COLING 2000*, pp. 90–96.
- Bresnan, J. (2001). *Optimal syntax*, pp. 334–385. Oxford University Press.
- Butt, M., Dyvik, H., King, T. H., Masuichi, H., and Rohrer, C. (2002). “The Parallel Grammar Project.” In *COLING 2002 Workshop on Grammar Engineering and Evaluation*, pp. 1–7.
- Crouch, R. S. (2005). “Packed rewriting for mapping semantics to KR.” In *Sixth International Workshop on Computational Semantics(IWCS-6)*.
- Fang, J. and King, T. H. (2007). “An LFG Chinese Grammar for Machine Use.” In *Grammar Engineering Across Frameworks Workshop*, pp. 144–160.
- Flickinger, D. and Bond, F. (2003). “A Two-Rule Analysis of Measure Noun Phrases.” In *HPSG03*.
- 飯田朝子 (2004). 数え方の辞典. 小学館.
- 見坊豪紀 (1965). “現代の助数詞.” 言語生活, **166**, pp. 54–60.
- King, T. H., Crouch, R., Riezler, S., Dalrymple, M., and Kaplan, R. M. (2003). “The PARC 700 Dependency Bank.” In *4th International Workshop on Linguistically Interpreted Corpora*.
- 荻野紫穂 (1993). “体言と副詞.” 星野和子, 丸山直子 (編), 日本語の表現, pp. 87–99. 圭文社.
- 荻野綱男 (1989). “現代日本語の助数詞の意味変化の方向.” 文藝言語研究 言語篇, **17**, pp. 67–77.
- 国語学会 (編) (1980). 国語学大辞典. 東京堂出版.
- 工藤拓, 松本祐治 (2002). “チャンキングの段階運用による日本語係り受け解析.” 情報処理学会論文誌, **43** (6), pp. 1834–1842.
- 黒橋禎夫 (2000). “結構やるな、KNP.” 情報処理, **41** (10), pp. 1215–1220.
- 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), pp. 79–109.
- Matsumoto, Y. (1993). “Japanese numerical classifiers:a study of semantic categories and lexical organization.” *Linguistics*, **31**, pp. 667–713.

- Maxwell, J. T. and Kaplan, R. M. (1993). “The interface between phrasal and functional constraints.” *Computational Linguistics*, **19**, pp. 571–589.
- 森重敏 (1958). “数詞とその語尾としての助数詞.” 国語国文, **27** (12), pp. 12–33.
- 日本電子化辞書研究所 (1996). EDR 電子化辞書説明仕様書.
- 岡部嘉幸 (2006). “数詞・助数詞に残る謎は何か.” 国文学, **51** (4), pp. 106–109.
- Riezler, S., King, T. H., Kaplan, R. M., Crouch, R., Maxwell, J. T. M., and Johnson, M. (2002). “Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques.” In *the 40th Annual Meeting of the Association for Computational Linguistics(ACL)*.
- 白井清昭, 徳永健伸 (2007). “呼応する名詞の包含関係に着目した助数詞オントロジーの自動構築と評価.” In *2007-NL-181*.
- 谷原公男, 顔瑞珍, DebbieLee (1990). “助数詞の用法とプロトタイプ—〈面〉・〈枚〉・〈本〉・〈個〉・〈つ〉.” 計量国語学, **27** (12), pp. 209–226.
- 特許 (2007). “自然言語処理システム及び自然言語処理方法、並びにコンピュータ・プログラム.” 登録番号第 4033093 号.
- Umemoto, H. (2005). “Implementing a Japanese Semantic Parse Based on Glue Approach.” In *20th Pacific Asia Conference on Language, Information and Computation (PACLIC20)*.
- 梅基宏, 杉原大悟, 大熊智子, 増市博 (2008). “LFG 解析と語彙資源を利用した日本語含意関係判定.” In *2008-NL-188*.
- 渡辺実 (1952). “日華両語の数詞の機能—助数詞と単位名.” 国語国文, **21** (1), pp. 97–109.
- 矢崎祥子 (1999). “助数詞の類別機能および数詞と助数詞の組み合わせ—和語・漢語の視点からみた六つ・六房・六個.” 言語と交流研究会, **2**, pp. 127–139.

略歴

大熊 智子 (正会員) : 1994 年東京女子大学文理学部日本文学科卒業. 1996 年慶應義塾大学政策・メディア研究科修了. 同年, 富士ゼロックス (株) 総合研究所入社. 2006 年より慶應義塾大学政策・メディア研究科後期博士課程在籍. 2009 年より東京女子大学非常勤講師. 現在, 富士ゼロックス (株) 研究技術開発本部研究副主任.

梅基 宏 (正会員) : 1995 年東京大学大学院工学系研究科機械情報工学修士課程修了. 同年, 富士ゼロックス (株) 入社. 日本語意味解析, 情報検索システムの研究開発に従事. 2004~2006 年 Stanford 大学 CSLI 客員研究員. 言語処理学会, 情報処理学会各会員.

三浦 康秀：2004 年電気通信大学大学院電気通信学研究科電子情報学専攻修了。同年、富士ゼロックス（株）入社。日本語の統計的構文解析、医療テキストを対象とした自然言語処理の研究開発に従事。

増市 博（正会員）：1989 年京都大学工学部機械工学科卒。1991 年京都大学工学研究科精密工学専攻修士課程修了。同年、富士ゼロックス（株）入社。1998～2000 年米国 Stanford 大学 CSLI 客員研究員および Palo Alto Research Center Inc. コンサルタント研究員。現在、富士ゼロックス（株）研究技術開発本部研究主査。博士（工学）。

（2008 年 12 月 12 日 受付）

（2009 年 1 月 28 日 再受付）

（2009 年 2 月 24 日 採録）