解　説

文法記述に基づく言語解析の進展と多言語処理の可能性

増市　博　富士ゼロックス（株）研究本部
hiroshi.masuichi@fujixerox.co.jp

大熊　智子　富士ゼロックス（株）研究本部
ohkuma.tomoko@fujixerox.co.jp

原田　康也　早稲田大学法学部・ことばの科学研究所
harada@waseda.jp

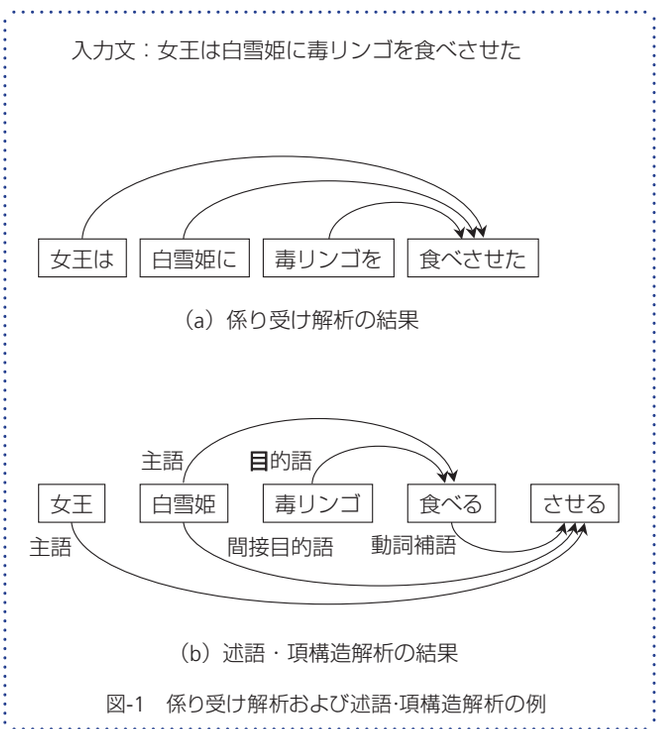
自然言語処理の分野では，1990年代以降，大量のテキスト情報を利用した統計処理手法に基づく言語解析の研究が大きな成果を収めている．一方，2000年代に入って，統計処理のみに基づく解析の限界も指摘され[1), 2)]，宣言的文法記述に基づく『深い』言語解析手法と統計処理手法の融合の重要性があらためて注目されている[3)]．本稿では，深い言語解析技術の理論・方法論・応用の可能性を現状の研究の進展状況とともに説明する．さらに，深い言語解析技術の汎用性のテストベッドとして，『複数の言語』を対象に深い言語解析技術の研究を行っているプロジェクトを概観し，その目標・成果等を具体的な事例を挙げながら述べることにする．

深い言語解析とそのアプリケーション例

『深い』言語解析といってもその解釈・定義はさまざまである．本稿では，文の構成要素間の修飾関係（係り受け関係）だけではなく，一般に『述語・項構造』と呼ばれる情報（文の中で，どの要素が述語であり，その主語・目的語は何であるかといった情報）まで特定する処理を深い言語解析と呼ぶことにする（図-1参照）．

このような深い解析を実現するには，解析の対象となる言語の文法規則をあらかじめ宣言的に記述しておくことが必要である．言語学的に精緻であり，実文を対象として高いカバー範囲を持つ計算機可読の言語文法を開発することは，言語の統語的性質を明らかにするという理論的興味からだけでなく，機械翻訳，文章生成，文書要約，文章自動校正，文章内容理解，質問応答といった自然言語処理アプリケーションの精度向上を実現するためにも重要であると考えられる．たとえば，質問応答システムにおいて，図-1の入力文から『誰が毒りんごを食べましたか？』の質問に正しく回答するための情報を得るには，図-1（b）の述語・項構造（すなわち『食べる』の主語が『白雪姫』であり目的語が『毒リンゴ』である構造）を特定する必要があり，深い解析が必須である．

後で述べるParallel Grammarプロジェクトでは，実験段階ではあるが，このような深い解析を利用して，質問応答システム等のさまざまな言語処理アプリケーションを構築している．本章では，それらの中から米Xerox社のEurekaシステム[4)]と呼ばれる文書管理システムの例を紹介する．



（a）係り受け解析の結果

（b）述語・項構造解析の結果

図-1　係り受け解析および述語・項構造解析の例

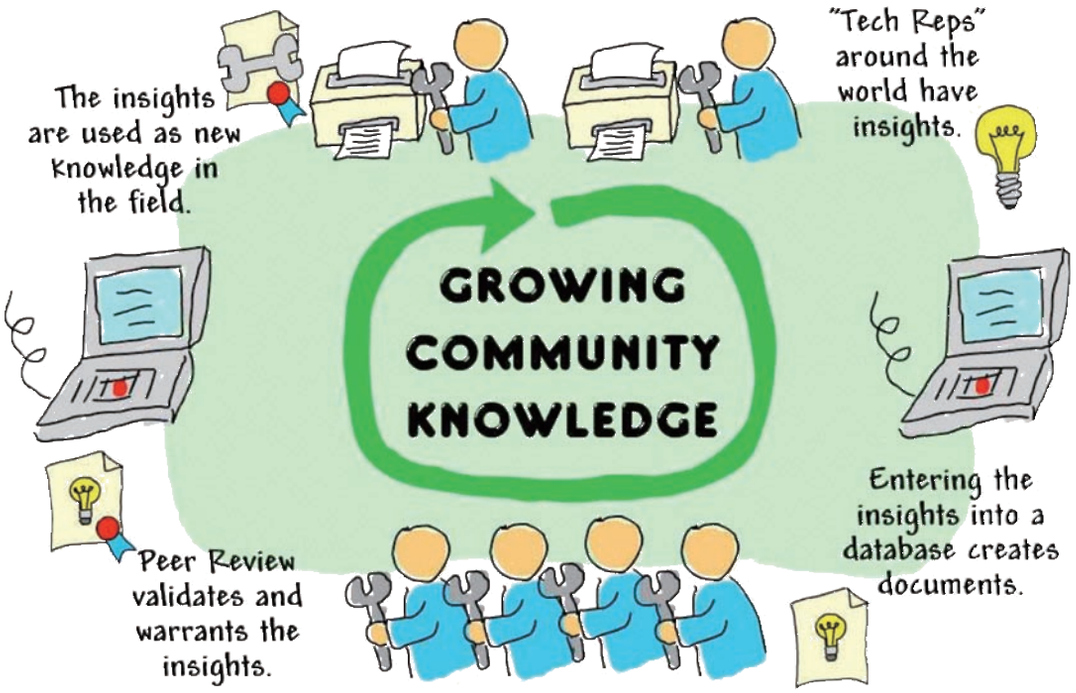


図-2　Eurekaシステムの概念図

Eurekaは一言でいえばノウハウデータベースである．顧客先に出向いてコピー機やプリンタ等の修理を行うカスタマーエンジニアが，自らの経験から得た修理のノウハウ（『問題点』とその『原因』・『解決策』）を記述し，データベースに登録する．登録されたノウハウを他のカスタマーエンジニアが参照することにより，修理業務を効率的に行うことができる（図-2参照）．しかし，Eurekaには現在約5万件のノウハウが登録されており，矛盾する情報が登録される，あるいは，冗長な情報が数多く登録される，といった問題が顕在化している．この問題を解決するために，Parallel Grammarプロジェクトでは，新しいノウハウの登録時に，以下の点を自動的に判定するシステムを構築している．

(a) 新しく登録されるノウハウに関係するノウハウはデータベースに登録されているか？
(b) すでに登録されているならば，新しいノウハウが含む情報は登録済みの情報と同一の内容であるか？（同一であれば冗長なノウハウを登録すべきではない）
(c) 同一でなければ，解決策の代替案を記述したものか，より良い解決策を記述したものであるか，既知の問題点よりも解決が難しい問題点について記述したものであるか，あるいは，既存のノウハウとの間で矛盾をきたす内容であるか？

上記 (a) の処理については，統計処理手法に基づく情報検索モデルを用いている．しかし，(b) ～ (c) については，単語を単位とする統計処理手法によって内容の同一性・冗長性・矛盾の検出を実現することは困難である．Parallel Grammarプロジェクトでは，以下の手順で上記 (b)～ (c) の処理の実現を目指している[5]．

(1) Eurekaに登録されたノウハウテキストに深い言語解析を施す．
(2) 得られた述語・項構造を述語論理式に変換する．
(3) 省略された主語や目的語の補完（ゼロ代名詞照応解析）や複合名詞の解釈を追加する．
(4) シソーラス，オントロジー辞書を用いて推論可能な知識表現形式に変換する．
(5) Structure Mapping Engine[6]と呼ばれる手法を用いてノウハウの内容を比較する．

図-3に，ノウハウテキストの例および上記 (1) (3) (4) の各手順の出力結果例（抜粋）を示す．これまでのところ，処理精度に関する評価結果は報告されていないが，解析カバー率（解析可能な文の割合）に関しては，(1) の深い解析のレベルで100%，(4) の知識表現形式のレベルで79%の値が得られている．また，最も計算量の大きい深い解析処理に要する平均解析時間は1文あたり約2秒であり，単語数が41以上50以下の範囲の長文に対する平均解析時間は約16秒である（2GHz CPU, 2GB RAM）．Eurekaシステムの研究担当者は，4～6年

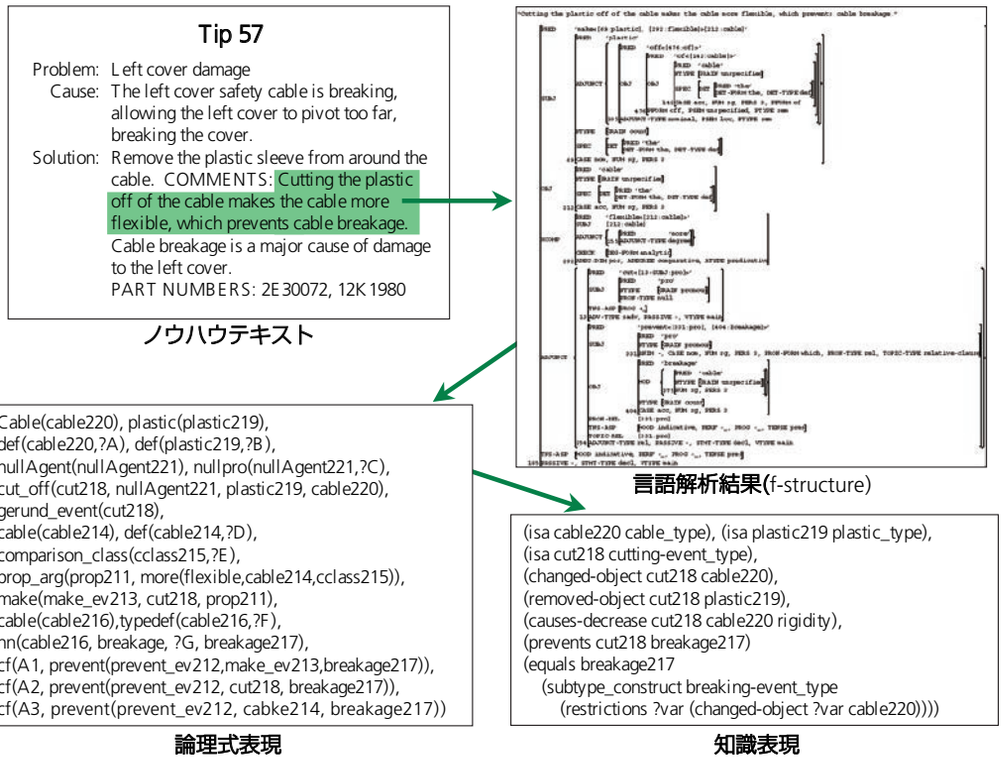


図-3　Eurekaシステムにおける処理過程例

のタイムスパンでこの技術を実用化することができる見込みであると述べている．Eurekaシステムで培われた技術は，将来的に特許文書，判例文書，医学文書といった価値の高い文書コレクションに適用することが可能であるとParallel Grammarプロジェクトでは考えている．Eurekaシステムは上記の通りさまざまな技術から構成されているが，深い言語解析技術はその中でも最も重要なコア技術として位置付けられている．

文法記述に基づく言語解析の理論と文法記述の実際

●深い言語解析の文法理論

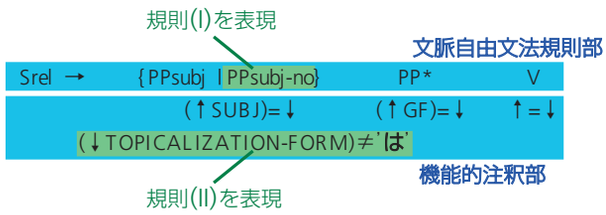
本節では，深い言語解析を行うための文法理論の代表例としてHPSG（Head-driven Phrase Structure Grammar：主辞駆動句構造文法）とLFG（Lexical Functional Grammar：語彙機能文法）を取り上げ概観する．

計算機または人間の認知的機構による『処理』を真剣に考えた場合，生成文法が広く採用する『変形』の概念をそのまま実装することは容易ではない．1970年代になってRichard Montagueの意味論が言語学者の広く知るところとなり，『心理的実在性』への関心から，『変形』を仮定しない文法理論が提案されるようになった．その成果の1つが後述するLFGであり，もう1つがHPSGである．

1970年代末にEssex大学のGerald Gazdarによって，それまで変形を用いなければ説明できないとされてきた言語現象（非有界依存や等位構造など）を，統語範疇を素性と値の組合せによって表現することにより，句構造文法によって簡潔に記述できることが示された．このアプローチは1980年代に入ってGPSG（Generalized Phrase Structure Grammar：一般化句構造文法）と呼ばれるようになる．HPSGは1980年代半ばにStanford大学CSLI（Center for the Study of Language and Information）において，計算機処理を前提としながら，GPSGに状況意味論やLFGの考え方を取り込みながら発展してきた．HPSGとLFGに共通に見られる特徴は文法を制約充足と見る点と厳密な語彙主義である．一方，近年のHPSGの特徴としては，『記号』に基づくアーキテクチャ，タイプ・タイプ階層・制約継承による言語情報の組織化，豊富な語彙情報からの一般原理による句の投射などを挙げることができる[7]．

HPSGの研究は当初から計算機実装と密接に連関していた．1980年代を通じてHewlett-Packard研究所が支援を続けたことがその名称に反映している．同じ時期，日本のICOTではGPSG/HPSGと同じような方向性で日本語の文法を理論的に厳密に記述する試みがJPSG（Japanese Phrase Structure Grammar）の名称で進められていた．

LFGは，1978年に現Stanford大学教授のJoan Bresnanによってその概念的枠組みが提唱された後，1982



Srel:関係節
PPsubj-no:格助詞「の」を伴い主格となり得る後置詞句
PPsubj:主格となり得る後置詞句(PPsubj-noを含まない)
TOPICALIZATION-FORM:主題化を行う係助詞の表層形

図-4　LFGに基づく文法記述例

年にXerox Palo Alto Research Center（現Palo Alto Research Center Inc.）のRonald Kaplanと，Joan Bresnanの共同研究によって定式化が完成した．LFGは，文に対してc(onstituent)-structureとf(unctional)-structureと呼ばれる2種の構造を措定する．c-structureは文の係り受け構造の木構造表現であり，一般に構文木と呼ばれるものに対応する．一方，f-structureは，主語（SUBJ）や目的語（OBJ）といった文法機能の概念に基づき，文の述語・項構造，時制，様相，話法等の意味情報を属性–属性値のリスト構造で表現するものである．LFGでは，c-structureを生成するための文脈自由文法（context free grammar）規則と，f-structureを生成するために文脈自由文法規則に付与する機能的注釈（functional annotation）を同時に記述する．

●LFGに基づく文法記述例

本節では，LFGに基づく文法記述の具体例を示す．この例のようにLFGでは，文脈自由文法規則に文法機能（あるいは他の素性）にかかわる制約を機能的注釈として付することによって，多様な文法現象を形式的に記述することが可能である．

(1)太郎が読んだ本
(2)太郎の読んだ本
(3)太郎の本を読んだ．
(4)太郎が本を読んだ．
(5)太郎は本を読んだ．
(6)太郎は読んだ本を捨てた．

たとえば，(1)(2)の名詞句は共に文法に則った表現であり『読む』の主語は『太郎』である．すなわち(2)においては格助詞『の』が主格標識として機能する．しかし，(3)の文において『太郎』が『読む』の主語であるという解釈は成り立たず，この場合の『の』は主格標識となり得ない．同様に，(4)(5)の文において『読む』の主語は『太郎』である．(5)では，係助詞『は』の主題化機能により主格標識『が』が省略されたと解釈できる．一方で，『太郎は読んだ本』という名詞句表現が不自然であることから，(6)の『太郎』は『捨てた』に係ると判断できる．すなわち，この場合『読む』の主語は省略されていると解釈する．これらの言語現象は，(I)『関係節内においてのみ格助詞『の』が主格標識として機能する』，(II)『関係節内において係助詞『は』による主題化は生起しない☆1』という文法規則に一般化できる．図-4に，LFG理論に基づいて(I)(II)の文法規則を形式化した例を示す．このように，深い解析を行うための文法記述は，言語現象を一般化して得られた文法規則を（既存の文法規則と矛盾しないかたちで）形式化する，というプロセスを繰り返し行うことが基本的な作業となる☆2．

複数言語を対象とする言語解析システム構築プロジェクト例

本章では，複数の言語を対象として並行的に深い言語解析システムの構築あるいはシステム構築のための文法記述を行うプロジェクトを紹介する．

欧州では1994年から1996年まで，LRE（Linguistic Research and Engineering）プログラムの下で，LS-GRAMと呼ばれるプロジェクトが複数の言語を対象とする研究活動を行った．LS-GRAMの対象言語は英語，ドイツ語，フランス語，スペイン語，ポルトガル語，デンマーク語，オランダ語，イタリア語，ギリシャ語の9カ国語である．ANLP（Advanced Language Engineering Platform）と呼ばれる共通のシステム上で単一化文法ベースの文法記述が行われた．ただし，後述の他のプロジェクトと比較して文法のカバー率は低いとの報告がなされている．

ドイツFederal Ministry of Educationの長期プロジェクトであるVerbmobilでは，PAGEとLKBという2つのシステム上で，ドイツ語，英語，日本語のHPSG文法の記述が行われた．Verbmobilで開発されたHPSG文法は，比較的短い口語的発話文を主な解析対象としている点に特徴がある．

Stanford大学CSLIが中心となって研究が行われているLinGO Grammar Matrixでは，主に英語，ドイツ語，日本語，スペイン語の文法開発を行っている．LinGO Grammar Matrixの英語文法はERG（English Resource Grammar）と呼ばれ，Verbmobilプロジェクトの機械翻訳システムでも利用された実績を持つ．LinGO

☆1　ここでは対比の『は』についての議論は省略する．
☆2　文法規則のほかに，単語あるいは形態素のレベルで言語現象を説明付ける語彙規則の記述も同様に重要な作業である．

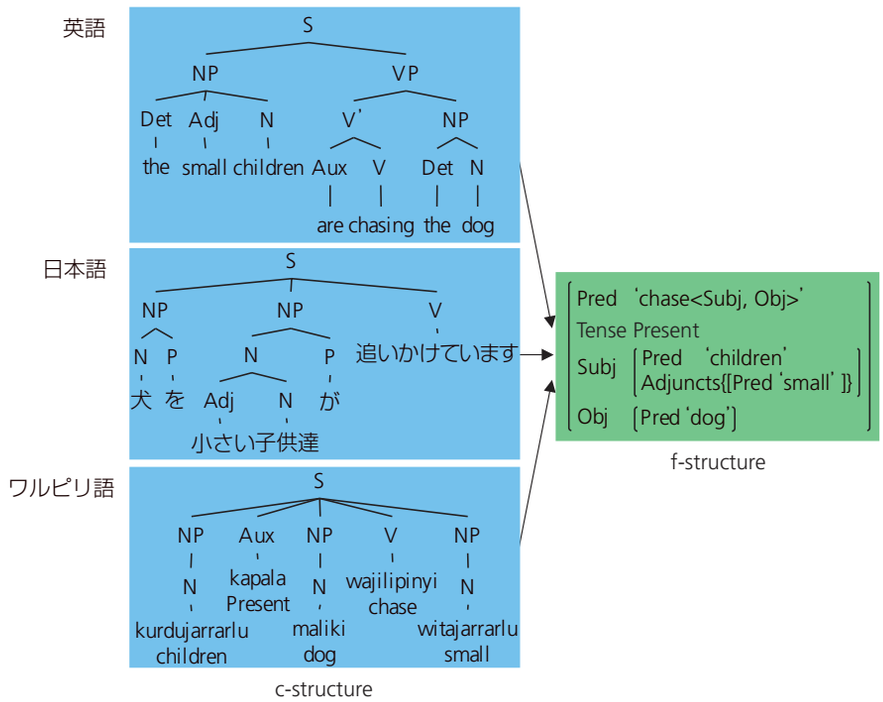


図-5　異なる言語を対象としたc-structureおよびf-structureの例

Grammar Matrixはオープンソースの形態で研究が行われ，比較的自由に新しい言語の文法記述を開始することが可能である．LinGO Grammar Matrixでは，MRS（Minimal Recursion Semantics）と呼ばれるフォーマットを用いることによって，さまざまな言語の解析結果をできる限り統一的に表現しようという試みがなされている☆3．

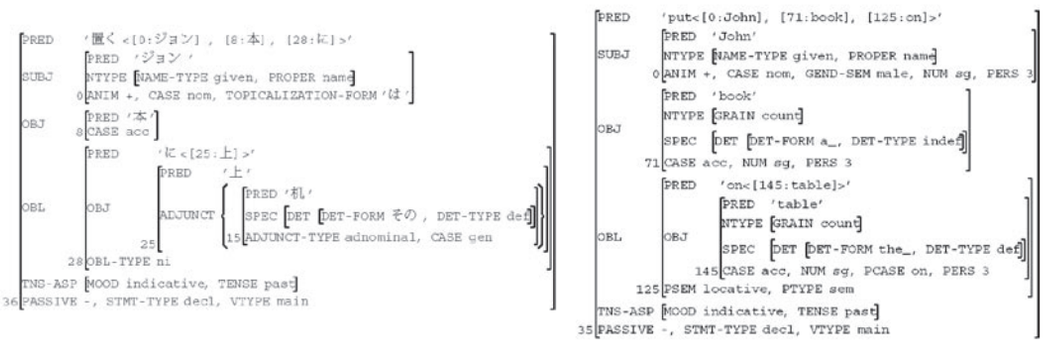
以下本章では，筆者が所属するParallel Grammarプロジェクト[8]について述べる．LFGの提案者の1人であるPalo Alto Research Center Inc.のRonald Kaplanがリーダを務めるParallel Grammarプロジェクトは，1994年に活動を開始した言語学者と計算機科学者によって構成される国際プロジェクトである．Ronald Kaplanのグループでは，1980年代後半からすでにLFGに基づく解析を計算機上で実行するための処理系開発を行っている．この処理系はXLE（Xerox Linguistic Environment）と呼ばれ，現在に至るまで高速化アルゴリズムの研究，開発を徹底的に行った結果，LFGの処理系としては世界最速の解析速度が実現されている．XLEの実装によって，これまで困難であるとされてきた大規模LFG文法に基づく実文の解析が初めて可能となった．

Parallel Grammarプロジェクトは，LFGとXLEを共通の基盤として，英語，ドイツ語，フランス語，日本語，ノルウェー語，ウルドゥー語，デンマーク語，韓国語を対象に，大規模文法記述を行うとともに言語解析システムの構築を進めている．Parallel Grammarプロジェクトの現在の主な参加組織は，Palo Alto Research Center Inc.（アメリカ），早稲田大学（日本），富士ゼロックス（日本），Stuttgart大学（ドイツ），DFKI（ドイツ），Bergen大学（ノルウェー），Constanz大学（イギリス），King's College London（イギリス），Dublin City大学（アイルランド），CLT（デンマーク）である．

LFGでは，文法機能をいかなる言語の解析においても有効であると考え，第一義の未定義要素としての地位を与えている．すなわち，見かけの表現がまったく異なっていたとしても，文法機能の概念はあらゆる言語において共通に存在するという立場をとる．言語が異なれば同じ意味内容を表現する文であってもその句構造すなわちc-structureは大きく異なる一方で，文法機能に基づく構造であるf-structureの違いは多くの場合きわめて小さいことが知られている（図-5参照）．言語処理システム構築の観点からは，解析システムの出力結果が言語を問わず一定であればあるほど，複数言語を対象とする言語処理システムの構築に要するコストは低減する．すなわち，単一のシステム構成で複数の言語を対象に同等の処理を行うことが可能となる．Parallel Grammarプロジェクトの目標は，冒頭で挙げたような自然言語処理アプリケーションを，言語の違いを超えて，統一的なシステム構成で実現することである．

Parallel Grammarプロジェクトでは，このf-structureの言語普遍的特徴を最大化するために，半年に一度全言

☆3　この試みは，後述のParallel Grammarプロジェクトにおけるf-structureの言語普遍性を最大化する方針と近いものである．



「ジョンは本をそのテーブルに置いた.」に対応するf-structure

「John put a book on the table.」に対応するf-structure

図-6　日本語文および英語文を対象としたf-structureの例

語の文法記述担当者が集まってミーティングを開催し，複数言語間でf-structureの整合性を可能な限り高める，あるいは，矛盾を回避する機会を継続して持つことにしている．すなわち，f-structureの構成から属性・属性値の用法やネーミングコンベンションに至るまでの詳細を議論し，標準仕様を決定する作業を行っている．**図-6**に，同じ意味を表す日本語文と英語文のXLEによる解析結果（f-structure）の例を示す．両者が基本的に等しい構造を持っていることが分かる．

Parallel Grammar プロジェクトで開発されたLFG文法のうち，英語，ドイツ語，日本語の文法は実用化に近いレベルに達している．英語解析システムは，Wall Street Journal誌の40,000文を対象として100%のカバー率を，日本語解析システムでもEDR（日本電子化辞書研究所）コーパス文10,000文を対象として98.3%のカバー率を達成しており，解析精度も各言語において高い水準にある[9), 10)]．

今後の課題

本稿で述べた通り，深い言語解析の技術は解析カバー率および精度の点においては複数の言語ですでに実用化のレベルに近づいてきている．しかしながら，原理的に計算量の大きい処理であることから，解析速度の問題は解決されていない[☆4]．Eurekaシステムの例のように，前処理として統計処理手法を用い，その後，必要な部分に深い解析処理を適用することが1つの解決策であると考えられる．また，文法規則では解消することができない解析結果の曖昧性を取り扱うためにも，統計処理手法との融合が今後の重要な課題であるといえる．

☆4 ただし解析速度の問題も徐々に解決される傾向にはあり，たとえば Parallel Grammar プロジェクトの英語 LFG システムの解析時間は Collins の構文解析システムと比較して 1.5 倍程度にすぎないとの報告もある[11)]．

謝辞　Parallel Grammarプロジェクトのメンバ，特に本稿の執筆に際して有意義なコメントをいただいたPalo Alto Research Center Inc.のRonald Kaplan氏，Tracy Holloway King氏，King's College LondonのMary Dalrymple氏に感謝いたします．

参考文献

1) 池原　悟：自然言語処理の基本問題への挑戦，人工知能学会誌，Vol.16, No.3, pp. 422-430 (2001).
2) 大谷　朗，宮田高志，松本裕治：HPSGにもとづく日本語文法について－実装に向けての精緻化・拡張－，自然言語処理，Vol.7, No.5, pp.19-49 (2000).
3) 郡司隆男，西田豊明，梶　博行，田中穂積：「言語理解－SHRDLUの先にあるもの－」へのコメントと回答，人工知能学会誌，Vol.15, No.6, pp.1035-1042 (2000).
4) Everett, J. O., Bobrow, D. G., Stolle, R., Crouch, R., de Paiva, V., Condoravdi, C., van den Berg, M. and Polanyi, L.: Making Ontologies Work for Resolving Redundancies Across Documents, Communications of the ACM, Vol.45, No.2, pp.55-60 (2002).
5) Crouch, R., Condoravdi, C., Stolle, R., King, T., de Paiva, V., Everett, J. O. and Bobrow, D. G.: Scalability of Redundancy Detection in Focused Document Collections, In Proceedings First International Workshop on Scalable Natural Language Understanding (ScaNaLU-2002), Heidelberg, Germany (May 2002).
6) Forbus, K. D., Falkenhainer, B. and Gentner, D.: The Structure Mapping Engine: Algorithm and Examples, Artificial Intelligence, Vol.41, No.1, pp.1-63 (1989).
7) アイバン・A・サグ，トマス・ワッソー著，郡司隆男，原田康也訳：統語論入門（上）・（下），岩波書店 (2001).
8) Butt, M., Dyvik, H., King, T. H., Masuichi, H. and Rohrer, C.: The Parallel Grammar Project, In Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation, Taiwan (2002).
9) Riezler, S., King, T. H., Kaplan, R. M., Crouch, R., Maxwell III, J. T. and Johnson, M.: Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques, In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadephia, PA (2002).
10) 増市　博，大熊智子：Lexical Functional Grammarに基づく実用的な日本語解析システムの構築，自然言語処理，Vol.10, No.2, pp.79-109 (2003).
11) Kaplan, R. M., Riezler, S., King, T. H., Maxwell III, J. T., Vasserman, A. and Crouch, R.: Speed and Accuracy in Shallow and Deep Stochastic Parsing, In Proceedings of the Human Language Technology Conference and the 4th Annual Meeting of the North American chapter of the Association for Computational Linguistics (HLT-NAACL'04), Boston, MA (2004).

（平成16年5月19日受付）