TM-0414

談話理解実験システム第3版と汎用日本語処理系の研究，開発構想

吉川貞行，内田俊一

November, 1987

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03) 456-3191～5
Telex ICOT J32964

１．はじめに

第５世代コンピュータプロジェクトは，通産省が推進している１０年計画のプロジェクトであり，１９８２年にスタートした．本プロジェクトの狙いを一言でいえば，「１９９０年代において最も重要なコンピュータ技術となる知識情報処理システムの実現にむけ，推論マシンの実現を核とし，コンピュータ技術の新しい体系（枠組）を構築すること」とすることができる．それはハードウェアとソフトウェアの両者の役割分担を見直し，ハードウェアの機能を推論をベースとする，より高度なものとし，それにより，より知的な機能を実現するソフトウェアの作成を可能にしようとするものである．本プロジェクトでは，ソフトウェアとハードウェアの境界線として，二段論法のような論理的推論をベースとする論理型言語を採用し，この言語を並列処理によって高速に実行する並列推論マシンの実現を目指している．

本プロジェクトでは，ソフトウェア研究の一つの大きな柱として，自然言語処理の研究・開発を取り上げている．それは，一つには人間にとって最も自然な自然言語による対話機能をマン・マシン・インタフェースとして実現することが目的であり，もう一つには，自然言語処理のシステムの実現には，推論機構や知識ベース機能，知識の表現や管理の方法，高速かつ大容量の処理を可能とする並列ハードウェアなど，諸々の機能要素が十分に機能することが前提となるからである．

すなわち，自然言語処理システムは，第５世代コンピュータを構成する各要素技術の実現度を反映する総合的なシステムとして位置付けることができる．

上記で述べたソフトウェアやハードウェアの基盤となる論理型言語は，もともと自然言語処理の研究を目的として生み出された言語である．本プロジェクトの自然言語処理研究の特徴の一つがここに有る．すなわち，自然言語処理の研究から生み出された論理型言語をベースに，計算機科学と言語学とを結び付けた高度な自然言語処理の技術を確立しようとしているわけである．

具体的には，談話理解を対象として各種自然言語処理の研究を進めながら，これと並行して，その枠組みの妥当性を確認するための談話理解実験システム(ＤＵＡＬＳ：Discourse understanding aimed at logic-based systems)の構築を進めている．また，ここで得られた技術をもとに自然言語処理研究・開発のためのツールとして位置付けられる汎用日本語処理系(ＬＴＢ :Language Tool Box)の整備を進めているほか，後期に向け自然言語処理の並列化についても基礎的な研究を進めている．

本テクニカル・レポートでは本プロジェクトにおける自然言語処理研究・開発の活動状況，談話理解実験システム第３版及び汎用日本語処理系の研究・開発構想について述べる．

２．研究・開発計画

本プロジェクトでは，人間と機械のインタフェースとして，従来のコンピュータの都合にあわせたコマンド言語ではなく，人間にとって都合のよい自然言語を用いることを目指し，自然言語処理の研究・開発を知的インタフェースという枠組みの中で行っている．

本プロジェクトの自然言語処理の活動を大別すると次の３つに整理できる．

①談話理解の研究・開発
②汎用日本語処理系の開発・整備
③自然言語処理の並列化の研究

自然言語理解の研究としては，現在，小学校の国語の文章を対象に談話理解の研究を進めている。これは，談話を記述した文章をコンピュータに読みこませ，各種自然言語処理を行うことにより理解させ，その内容に関する質問に答えられるようにすること目論んでいる。この実験用のコンピュータには，本プロジェクトで開発した論理型プログラミング用のワーク・ステーション（ＰＳＩ）を使用している。

また，これらの実験を効率的に行うためには，自然言語を扱うための種々のソフトウェア群が必要になる。このなかには，エディタやデバガなどの開発環境や形態素解析，構文・意味解析などの自然言語処理に共通的に用い得るソフトウェア・モジュールが含まれている。そこで，これらを汎用日本語処理系と言う共通ソフトウェア・モジュール群として構築している。

しかし，これら形態素解析や構文・意味解析自体にいまだ解決しなければならない課題も多く，また談話理解のための改良なども考えられる。そこで，追加，修正，変更ための機能について充分考慮し，充実させている。

３．談話理解の研究・開発

(1) 研究・開発の意義

談話理解は，現在進められている人工知能や知識情報処理の研究の中でも，最も高度な研究対象として位置付けられている。それは，「言語を理解する」とはどういうことか，という基本的な点についてばかりでなく，話者の意図，情緒など現在の技術では定式化できていない言語現象が多数存在するからである（これらは，文章の理解を踏まえた高品質な機械翻訳などの自然言語処理システムを実現するためにも解決を必要とするものである）。このため，談話理解は，基礎的，理論的な研究が必要不可欠な分野である。また，そればかりではなく，実験システムを構築し，その枠組みの妥当性を確認していく必要が有る分野でもある。このために我々は，意味，談話構造，各種言語現象等の基礎的，理論的な研究と並行して談話理解研究のための実験システムであるＤＵＡＬＳを開発している。この実験システムは，段階を踏みながら開発を進めており，すでにＤＵＡＬＳ－Ⅰ，Ⅱを開発している。

(2) 談話理解の基本構造

談話理解の基本的な構造は，図１のように考えることができる。図１を入力から順に概説する。

まず理解する対象である文章（漢字かな混じり文）が入力される。文解析部では，入力された文章を一文単位に区切る。そして一文単位に以下の処理を行っていく。まず，辞書引きを行いながら文を単語（品詞）単位に分解する形態素解析を行う。次にこれらの単語（品詞）列の並びが文法規則に当てはまるかどうか，係り受け（修飾関係）はどうなっているか，意味的におかしくないか，などの処理を行う構文・意味解析を行う。これらの解析が正常に終了すると，文解析部は，入力された文の単語の並び，係り受け関係，文法構造などをツリーの形で表現した構文解析木と意味をフレーム形式で表現した中間表現として出力する。これで，この一文の構文・意味的な構造が分かったことになる。

次にオブジェクト同定部が，この中間表現を参照し，主語が省略されている場合はその主語は何か，「これ」，「それ」，「あれ」などの指示代名詞があればそれは何を指しているか，などの省略，照応処理を行い，記述状況と呼ばれる一文の解釈を出力する。

この記述状況を談話解析部が受け取り，これまで蓄積した談話の情報（文脈）から，この文が単なる詠嘆／反問／自問の文なのか，それとも疑問文なのか，などの発話行為の解釈を行った後，談話構造として蓄積する。

これらの処理を入力された文章が終わるまで繰返し行う。

先ほど述べたように，談話解析部では，疑問文か否かの判定も行っている。疑問文の場合は，問題解決部にその記述状況を引継ぐ。問題解決部では，疑問文に対する解を談話構造の中から検索，あるいは知識を使って推論し求めている。

文生成部は，求めた解から日本語文の生成を行っている。

(3) 研究・開発の流れ

本プロジェクトの開始当初においては，談話を理解するシステムを開発するために必要な研究要素を探るため，これまでの言語学，構文解析などの計算機言語学，論理学，知識表現などの研究を一つの統一的なシステムとしてまとめみよう，というところから研究を始めた。

このために開発されたのがＤＵＡＬＳ－Ｉである。まず，構文解析として，構文解析木を下から上へ向かって作って行く方式である，ボトム・アップ・パーザ（ＢＵＰ）を，基本的な枠組みとなる意味構造として，スタンフォード大学の J.Barwise と J.Perry らによって探索された状況意味論 (Situation semantics)を採用した。

また，日本語の省略，照応の解析を行うオブジェクト同定処理には，亀山アルゴリズムを適用した。記述言語は，その当時まだエジンバラProlog しかなく，しかたなくこれを用いている。

このようにして１０文，１００単語程度の文章を読み，質問（あらかじめ準備した質問）に答える小規模な実験システムが出来上った。そして，この小規模な実験システムから多くの研究課題とアイデアが浮かび上がってきた。その一つが意味記述言語（ＣＩＬ）である。この言語の特徴等は後述する。

ＤＵＡＬＳ　Ⅱでは，これまでの経験をもとに体系化し，あるていど技術的に固まってきたものから汎用化を行った。まず第一に各機能モジュールに分割を行っている（図１を参照），第二にプログラムの記述をＣＩＬに統一し，そのアイデアが実際の談話理解に十分な力を発揮できるかどうかの確認を行った。第三に，文解析部の文法の充実，またより高速化するために，並列処理のアイデアを導入し，処理速度を向上を図っている（ただし，ＰＳＩ上では逐次的に動作），また，ＤＵＡＬＳ－Ｉでは解答文対応にアドホックに開発した文生成を，
中間表現から日本語文を生成する文生成部として新たに実装している。

文解析部，文生成部等の汎用化により，これまでの決った質問文だけでなく，各種質問も受付けられるよう改良を加えている。ただし，この段階では辞書の語彙数と問題解決部の一般的知識数の貧弱さのために受付けられない質問文も少なからずあったが，設計者も予測しなかったような質問に正しく答えられたこともあった。

４．談話理解実験システム第３版の研究・開発構想

現在，我々は６３年度末を目標にＤＵＡＬＳ－Ⅲ の研究・開発を手懸けている。これまでＤＵＡＬＳ－Ⅰ，Ⅱと１０文，１００単語程度の小規模の文章を対象に実験を行ってきたが，これらで得られた技術をもうすこし規模の大きい２００文，２，０００単語程度の文章に対し適用し，その妥当性の確認を行うことを計画している。

これまでは，実験規模が小さいこともあり，単語（語彙）の意味や省略，照応などの処理は，ある程度この実験用の文章に限定したものであった。しかし，実験に量的な拡

大や文章の質に変化を与えることにより，技術の質的な変化，進歩が求められるようになる．たとえば，単語（語彙）の意味記述では，その単語が現われる多くの文章で利用できるように，より精密に汎用的に記述しなければならないばかりか，その量的拡大により処理時間などの問題が生じてくる．

しかし，これらの問題を着実に解決して行くことによって，より広範な知識情報処理に適用可能な新しい技術が生れてくるのである．

ＤＵＡＬＳ－Ⅲの主な研究・開発課題としては，談話内容の時間的経過や意図とその関係などの構造化するための談話構造，省略照応処理を行うオブジェクト同定，単語（語義）の意味表現，談話の焦点を明らかにし管理する焦点処理，自然な日本語文を生成するための文生成プランニング，そのほか多くの談話における各種言語現象がある．

５．汎用日本語処理系の研究・開発構想

ＤＵＡＬＳ－Ⅲを実現するためには，これまでのＤＵＡＬＳ－Ⅰ，Ⅱのような小規模な実験からある程度の規模の研究実験ができる，研究実験環境を整備する必要がある．それで我々は当面の課題としてＤＵＡＬＳ－Ⅰ，Ⅱの経験をもとに言語データベース／辞書，意味記述言語，形態素解析，構文・意味解析，文生成について整備を進めている．

また，現在基礎研究として進めている，より高度な知識の表現手法や推論の手法，シソーラスや概念辞書の構造などについても，今後その具体的な実現方法等が明確化した段階で，それらのソフトウェアを共通ツールとしてとりあげ，汎用日本語処理系の一部として付け加えていく方針である．そして，本プロジェクトの最終段階においては，この汎用処理系が実際に広く利用可能なソフトウェア・ライブラリの形で提供できる成果となることを目指している．

現在の活動状況は次のとおりである．

(1) 言語データベース／辞書

自然言語処理においては，辞書をしっかり作ることは非常に重要なことだと考えている．それは，形態素／構文／意味／文脈／生成などの全ての処理に関係する基本的な仕組がここに集約されるからである．

辞書に反映する情報を大きく分けると，形態的情報／構文的情報／意味論的情報の３つに分けられる．形態素情報は，語の形態に関する分類，発音など，構文的情報は，文法的な特徴に関する情報で，品詞や構文的制約などである．最後の意味論的情報は語の意味（語義）を表現するもので，文や文脈の意味を解析するのに必要となるものである．

今後の中心的研究課題である意味論的情報では，動詞に関しては深層格の記述を，名詞に関してはその概念を論理式やシーソラスなどで表現すべく記述実験を行っている．

(2) 意味記述言語と開発環境

意味記述言語（ＣＩＬ）は，自然言語処理システム記述，意味表現などに適用することを目的に開発した言語である．そのために，Prologを拡張し，自然言語処理に使われるフレーム，属性／属性値の対リストなどの各種データ構造を表現出来る部分項と，ある変数が決った時点で，ある特定の処理を実行する（デーモン）遅延実行制御を取り入れている．

現在，プログラミングの効率化を図るため，スクリーンデバガ，リスナなどの開発，

コンパイラの高速化等の開発環境の整備と，インヘリタンス機能の追加等言語機能の向上を進めている。

(3) 形態素／構文・意味解析

文解析部には，並列処理のアルゴリズムを導入した，形態素解析ツール（ＬＡＸ）を，構文・意味解析ツール（ＳＡＸ）の開発を行い，ほぼその開発を終了し，現在そのデバグ環境を整備している。

(4) 文生成

文生成は，構文・意味解析の出力する表現形式（一部異なっている）から日本語を生成するものである。一つの表現形式からは，一つの文（表層文）を生成している。
今後の課題としては，一つの表現形式から文生成プランニングにより複数の文（表層文）を生成することが上げられる。

６．並列化への研究

これまで述べてきたような，各種自然言語現象を扱い解釈していくためには，これまで以上の計算量が必要となってくる。たとえば，これまでの実験例では，複雑な文章を入力すると，場合によってはその文の解析だけで数十分以上もかかることがある。これでは，より高度な処理を行うことは時間的に不可能となる。つまり，より高度な自然言語処理を実現していくためには処理の高速化が必須となって来るわけである。
これを解決する一つの方法として，処理の並列化がある。我々は，本プロジェクトで開発した論理型並列言語（ＫＬ１）を用いて並列推論マシン（ＰＩＭ）上に上記ＬＡＸ，ＳＡＸなどの自然言語処理ツールを再構築することにより，並列処理による自然言語処理実験環境を実現したいと考え，現在その基礎的研究を進めている。

７．おわりに

これらＩＣＯＴの自然言語に関する研究は，意味まで踏まえたより高度な機械翻訳のベースとなるものである。談話理解自体は，すぐに実現できるわけではないが，ここで進められている研究は，機械翻訳システムやその他多くの自然言語処理システムに，その要素となる技術を提供していくことになる。
また，これまで述べてきたような問題に前向きに取り組んで行くことが今後の自然言語処理の健全な発展をはぐくんでいくことになると考える。

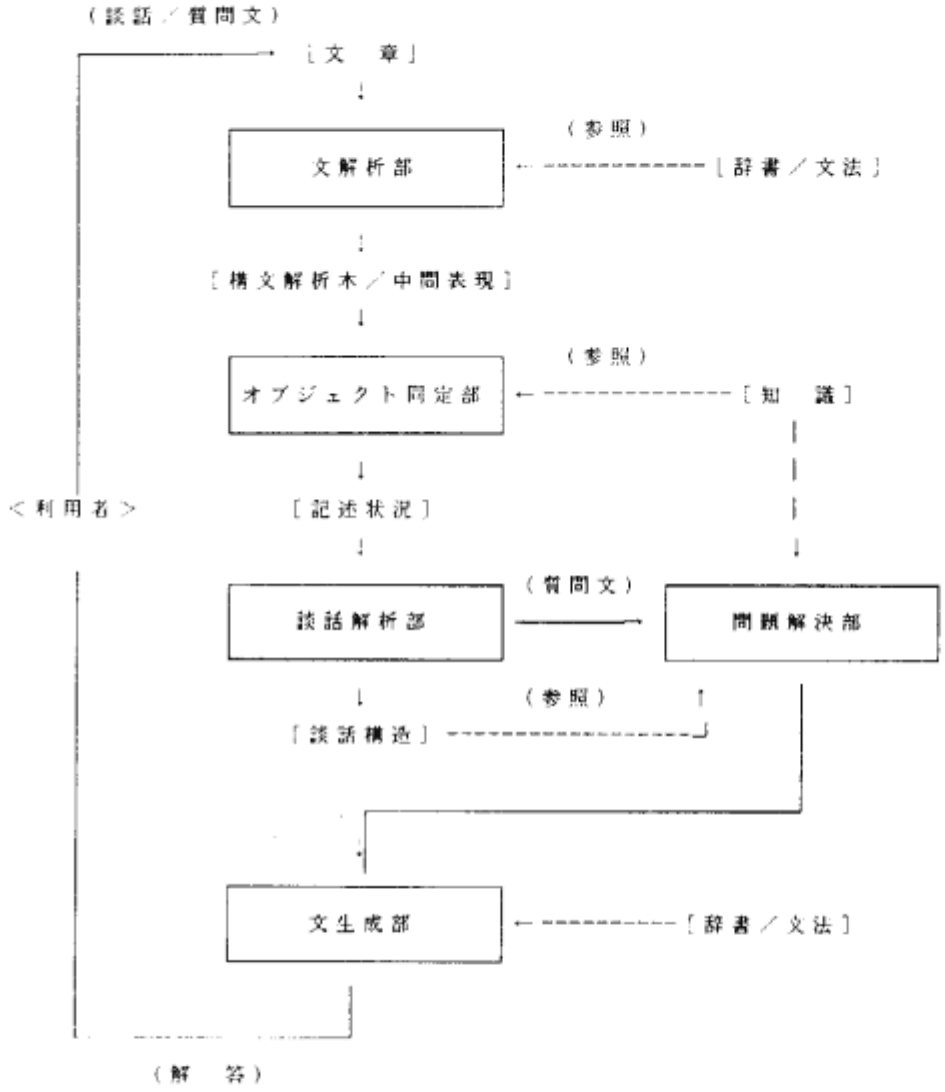

図１　談話理解の基本的な構造