自然言語処理 38－5
（1983. 7. 15）

日本語構文解析

辻井 潤一
（京大・工学部）

§1. はじめに

「入力文の持つ曖昧さをどのように取扱うか」が構文解析の最大の課題である。ある文脈に置かれた文が，人間にとってはほとんど一義に解釈できるように，文脈や「対象分野についてのプラグマティックな知識」を含めて，人間の持っている種々の知識と等量の知識をシステムが持っていれば，入力文は曖昧さなく解釈できるはずである。しかしながら，人工知能分野での自然語理解の研究が示すように，計算機が人間と同じように文章を理解することは，対象分野を極端に限定した世界においてさえ非常に困難である。現実的な機械翻訳システムは，これよりもはるかに広範な分野を対象としなければならない。特に，我々が本研究で対象とした文献アブストラクトでは，電気工学といった特定の分野に対象を限定したとしても，少なくとも数万語の語彙を必要とする。このような分野を対象とする機械翻訳システムの構文解析技術においては，一意に解釈を決定できないとすれば，「あらゆるレベルの情報を使って，正解の可能性が最も高い解析結果を最初に得る」ことが重要になる。文献(1),(2)で述べた，並列句のスコープを確度の高いヒューリスティック規則から順々に適用して決定する手法は，この端的な例になっている。

構文解析において，「意味」や「知識」が重要な役割を果していることは事実であるが，これにあまり強く依存することは，現実の機械翻訳システム用の構文解析としては危険である。従って，本システムでは，「意味」や「知識」処理が有効で，かつ，現実に行なえる場面ではこれを積極的に活用する枠組を用意する（open－ended）がそれ以外の様々な情報（語の形態的類似性，特定の単語固有の規則）を有効に使うことにより，「意味」や「知識」にもとづく処理が不十分であっても，正解の可能性が最も高い解析結果を出すことを目指している。

§2. 全体の処理の流れ

構文解析は，文献(3)の形態素解析が終了した段階のデータから出発して，トランスファ過程の入力となる文の意味構造を反映した依存構造表現を作り出す。構文解析の方針を以下に列挙する。

(1) 解析結果は，格文法に基く依存構造表現とする（文献(4)）。
(2) 語を構文的振舞いの観点から細分類し，キメの細い規則を設定する。（文献(5)）。
(3) 単語，特に名詞の意味を51の意味カテゴリーに分類し，動詞と名詞の共起制限を，動詞の格フレームに記述する（文献(5)）。
(4) 処理は，連用中止法・並列名詞句等，構文的なスコープ決定をまず行ない，曖昧さなく決定できる部分から順次構造決定を行なう。
(5) できるだけ多くのレベルの情報を参照しながら構造決定を行なってゆくために，中心となる木構造に，辞

※）本研究は，国の科学技術振興調整費による「日英科学技術文献の速報システムに関する研究」の一部として行ったものである。本研究は，長尾真・中村順一（京大），高松忍（阪府大），平井裕（豊技大），首藤裕（筑波大）の各先生，及び，隅田（日本IBM）・谷口（京セラ）・坂本（沖）・小坂（日電）・加藤（ソード）・石川（京大院生）・高井（JCS）の各氏との協同作業の結果をまとめたものである。

書情報・形態素情報・統語情報・意味情報等の種類の異なる情報を属性－属性値の形式で付与しておき，規則中で適宜これらを参照する。

(6) 慣用句および慣用句的表現を積極的に活用して曖昧さを解消してゆくために，GRADEの単語個別の辞書規則を使う[6]。

(7) 構文論的に可能な構造をすべて対等に取扱うのではなく，現実のテキストを綿密に調査することにより，「正解の可能性が最も高い解析結果」をできるだけ優先して出力する。このためのヒューリスティックな規則を積極的に取り入れる。

このような基本方針に従った構文解析の流れを図1に示す（図中，枠で囲んだ部分はGRADEの部分文法のネットワークで実現される）。

§3. 連用中止法と並列名詞句のスコープ決定

論文抄録のように，多くの情報を限られたスペースに簡潔に表現しようとするテキストにおいては，連用中止法や「と」・「及び」・「・」で結合された並列名詞句のような並列表現が多用される。このような並列表現がどの範囲の句を並列に結合していると判断するかは訳文の構造に直接あらわれる重要な問題である。

[例] 論理素子と回路の複雑性

→ { ・logic elements and complexity of circuits
  ・complexity of logic elements and circuits

並列名詞句の処理については，各種の表層上の手がかりや語の特殊性を活用するヒューリスティック規則で，かなり確度高く処理できることをすでに

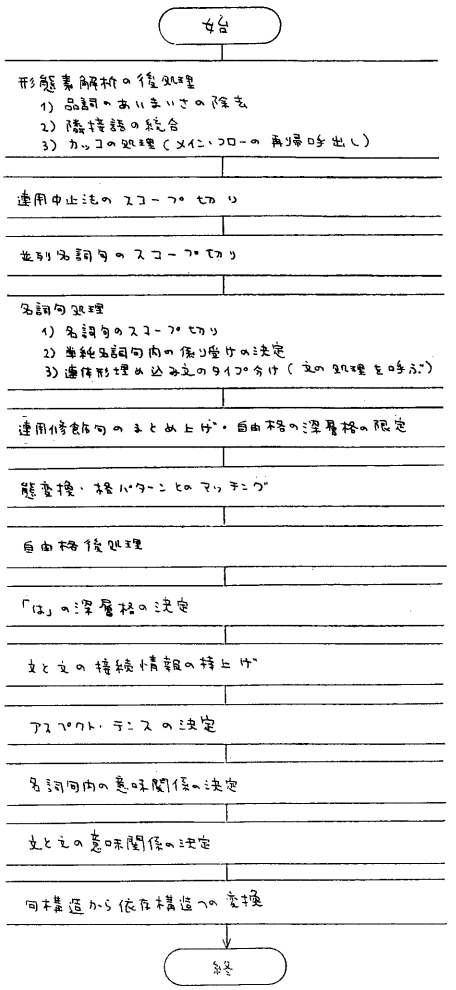


図1. 全体のフロー

報告した[1]。連用中止法がどの述部と並列になるかの決定は，述部の中心となっている動詞を次の4種類に分類することによって行われる。(*)

(1) 記述動詞：論文作成者が論文内容を説明するのに使っている動詞で，意志動詞であるもの（ex：述べる，報告する，記述する，実証する，現在約30語）

(*) ここで示す動詞の分類は，各動詞の抄録文中の用法に基づくものであり，一般分野のテキストにおいては，ここでの分類以外の用法を持つものもある。

(2) 準記述動詞：記述動詞と同様の使われ方がされる動詞で，無意志動詞のもの（あるいは，ひとまとまりで述部を作る複合表現）（ex：明らかになる，etc.）
(3) 意志動詞（ex：決定する，使用する，etc.）
(4) 意志・無意志動詞（ex：検出する，変える，etc.）
(5) 無意志動詞（ex：比例する，低下する，放出する，etc.）

ここで，記述動詞は，抄録文中では，文末の述部として用いられるか，あるいは，連用中止法の述部として現われた場合には，文末の述部と並列になることがほとんどであるような動詞である。例えば，記述動詞「述べる」は，

[例] ・・・理論について述べ，n-GaAs面を基板とし，アルミナ膜を絶縁膜としたMISキャパシタを，・・・・・測定を行い，結果を解析した。

のように，途中に構文的には並列になり得る述部があっても，それを飛び越して文末の述部と並列になっている。また，上記の例は，「～とし，～とした」のように同一の述部の繰返しがある場合には，その述部同志が並列になり易いというヒューリスティック規則の例にもなっている。現在，このような動詞の分類に基づく連用中止法のパターンとそのヒューリスティック規則を，現実のテキスト（1万文のデータベースを持っている）から分析し，それをGRADE規則として定式化する作業を行っている。人間によるテキスト分析の結果では，1000文中に256例の連用中止法を含む文があらわれ，現在の規則化で約200例については正しく並列の相手が認定できる。連用中止法については，この範囲推定の問題以外に，つながれた2つの文の間の意味的関係を同定する問題があり，質の良い訳文を生成するために，現在その検討を進めている。

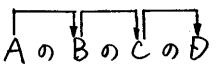
§4 単純名詞句の処理

（形容的連体詞を除く）連体詞と「名詞＋（連体助詞）」だけからなる系列を単純名詞句とよぶ。論文表題の場合と同様，論文抄録においても比較的長い単純名詞句があらわれる。このような系列は，

[例] 長さが5cmの棒
パラメータが未知の場合
本システムで考慮中の手法

の各下線部のように，単純名詞句の形式をしている場合であっても，述部の「である」が「の」に縮退していて，実際には単純名詞句として組み上げると誤解釈になる場合もあるが，上記のような例は，実際のテキスト中にはまれであり，前後の単語の意味的属性や単語そのものを参照すれば，縮退した述部「である」を復えできることが多い。このような復えの後，単純名詞句が処理される。

単純名詞句内での係受け関係は，系列中に動作名詞を含まない場合には，



のように，単純なものが多い。これに名詞の細分類や意味マーカを使った例外的なヒューリスティック規則，例えば，

・連体詞は固有名詞・記号を修飾しにくい
・限定連体詞は動作名詞を修飾しにくい
・連体助詞'についての'の後続した名詞句は，意味マーカ工丁の名詞を修飾する

ことが多い，etc.

を活用することにより，多くの場合，正しく係受け関係を同定できる。動作名詞が系列中に含まれている場合には，その動作名詞の格構造記述と名詞の意味素性の照合が行われる。

§5 連体形埋込み文の処理

日本語の連体形埋込み文は，次の4つのタイプに分類される。

(1) TypeⅠ：被修飾名詞が埋込み文中の欠けた格要素に対応するもの
(2) TypeⅡ：被修飾名詞が埋込み文の格要素となる名詞と意味的な関係を持つもの
(3) TypeⅢ：埋込み文と被修飾名詞との間に「～という」を挿入することができる同格表現
(4) TypeⅣ：埋込み文のあらわす事実と被修飾名詞とが何らかの意味的関係を持つ，いわゆる部分的同格関係の表現(7)

ここでは，埋込み文で表層上の「が」-，「を」-，「に」-格に入る名詞句が被修飾語として取り出されたような構造のみをTypeⅠと考え，それ以外の格（いわゆる，場所格・時間格・理由格等の自由格）に入るべき名詞句が取り出された構造は，TypeⅣに分類することにした。1000文のテキストにあらわれる埋込み文を，分類した結果を表1に示す。

タイプ	Ⅰ	Ⅱ	Ⅲ	Ⅳ
頻度	287	21	146	85

表1 埋込み文のタイプ

この表から判るように，TypeⅠの埋込み文が圧倒的に多い。この型の埋込み文は，動詞の格構造記述と被修飾名詞句の主名詞の意味マーカの照合によって処理される（この際，埋込み文の態や動詞・名詞の細分類が参照される）。他のタイプへの分類は，被修飾名詞や述部の素性等の，種々のレベルの情報を参照するヒューリスティック規則を使って行われる。例えば，次のようなヒューリスティック規則が考えられている。

(a) TypeⅡは，埋込み文中で「が」-格，「を」-格に入った名詞が，部分・属性・動作を表わす名詞で，埋込み文の述部が状態性述部である場合に考慮される。
(b) TypeⅢは，被修飾名詞が特定の名詞である場合に考慮される（ex：必要(性)，場合，こと，方法，傾向，現象，問題，etc. ・・・この種の名詞として，現在30程度のものが抽出されているが，これらを他の名詞からうまく分離できる意味マーカはない。したがって，この種の名詞は，別の属性を設定して，個別に辞書中にマークされる。また，この種の名詞は，「こと」のように完全な補文標識であって，TypeⅢだけを考慮すれば良い名詞と他のタイプの可能性も考慮する必要のある名詞とにさらに細分される）。
(c) TypeⅣに分類される埋込み文は，埋込み文のあらわす事実と被修飾名詞との意味的関係に従って，さらに9個のサブタイプ（現在）に分類されているが，個々のサブタイプに応じて，例えば，次のような表層上の手がかりを利用することができる。
(c-1) 被修飾名詞が特定の語基を持つ名詞であること（ex：Alの特徴を生かした応用例，スクリーンが集束能力を持つパラメータ範囲，etc.）

(C-2) 埋込み文が特定の述部を持つこと（ex：フォーカス装置を用いた高磁場，ケーブル材としてAL管を使った方法，etc.）

(C-3) 特定の格助詞相当表現が埋込み文と被修飾名詞との間に挿入されていること（ex：劣化を適当な程度で防ぐための指針，etc.）

以上に示したような手がかりは，あくまでヒューリスティック規則であって，その手がかりが存在すれば一意に構造が決定できる性質のものではない。したがって，ここでも並列名詞句のスコープ決定で行ったのと同様に，比較的安定した，成功率の高い規則から順々にヒューリスティック規則を適用してゆき，「正解の可能性が高い」タイプ決定を行ってゆく必要がある。

§6 テンス・アスペクトの処理

名言語は，文であらわされた事実が時間軸上でどの位置に位置するか，また，事実間の時間関係がどのようになっているかを表現するために，テンス・アスペクトの表層上の表現形式を持っている。この表層上の表現形式は各言語固有であり，とくに日・英間では絶対テンスと相対テンスの体系の相違があるために，この差を解析段階で調整しておくことは，トランスファ過程の負担を軽減する意味からも重要である。

アスペクトの解釈をするために，日本語の動詞を次の6種類に分類している。(*)

(1) 状態動詞：ある，いる
(2) 準状態動詞：交叉する，異なる，所属する
(3) 瞬間動詞：衝突する，起る
(4) 瞬間（結果）動詞：分離する，省く，判明する
(5) 継続動詞：飛行する，考慮する
(6) 継続（結果）動詞：成長する，広がる，改善する

(*) 動詞辞書の第一版を作成する時点では，分類の作業基準が不明確であったために，第一版辞書では，状態・瞬間・継続の3区分が記載されている。

ここで，準状態動詞は，論文抄録を処理するために特に設けた分類であり，アスペクト形式素「ている」がついてもほとんど意味が変化せず，論文抄録の範囲内では，φ－形・ている－形ともに状態を表現していると考えて良い一連の動詞である（また，この種の動詞は，連体形埋込み文中でタ－形であらわれても，テンス形式素「た」の意味はでず，超時的状態を示すことが多い）。

動詞の分類（アスペクト素性）とアスペクト形式素の組合せによる深層アスペクトの解釈規則の一部を表2に示す。

形式 / 素性	する	している	してくる	しつつある
準状態	状態	状態		近未来
瞬間	超時	反復	反復	近未来
瞬間結果	超時	結果	反復	近未来
継続	超時	継続		継続
継続結果	超時	推移	継続	継続

表2 アスペクトの解釈

この表に示された解釈は，他に深層のアスペクトを決定する手がかりがない場合に，動詞の素性とアスペクト形式素だけから解釈を決定するためのデフォルト規則である。したがって，文中の他の要素（副詞，副詞句・節，埋込み文の場合には，その埋込み文を導く補文標識的な名詞，etc.）によって，解釈は変更されることになる。

[例]・彼は本を読んでいる。
・彼はしばしば本を読んでいる。

副詞の辞書には，この目的のために，各副詞がどのようなアスペクト解釈を強要するかが指定されている。

日本語の相対テンスの体系を，英語の絶対テンスの体系に変換する処理は，主節のテンス決定から順次下位の補文に向って再帰的に行う基礎のアルゴリズムを基本的に採用している(8)。この過程も，主節・従属節の関係と述部の相対テンス形とだけを参照するデフォルト規則と，時間を示す副詞等の存在から解釈を決定する，より強い規則とに分けて行われる。

§7 動詞と名詞の共起関係

単文内の係受け関係の処理は，基本的には名詞句の主名詞と動詞の共起可能性を判定することによって行われる。文献(9)で論じたように，各動詞の辞書には，その動詞の用法に応じて，複数個の格構造記述が用意されており，単文内の係受け関係を処理する時点で動詞がどの用法で使われているかが判断される（動詞が複数個の'意味'を持っている場合には，それぞれの意味に応じて，異なった格構造記述が用意されている）。各格構造記述には，それに対応するトランスファ辞書中の項目が指定されているので，この解析の時点で，複数個の意味を持つ動詞の場合には，意味の分離が行われたことになる。厳密には，名詞－動詞の共起関係によって，名詞の側の意味の分離も行われることになる。

文献(9)でも指摘したように，名詞と動詞の共起関係には，置換可能性の高い共起関係と，置換可能性の低い慣用句（的）表現のような共起関係とがある。置換可能性の低い共起関係については，現在の名詞意味マーカ程度の荒い記述では対応できないので，共起する名詞を直接動詞辞書中に指定する方式をとっており，語基レベルでの照合で対処する方針である。準（あるいは準々）慣用句的表現に対処するためには，語基レベルでのシソーラスといったものを考える必要があるかもしれない。この種の共起関係は，名詞－動詞間だけでなく，副詞・形容詞（連用形）と動詞の間にもみられるので，これらも記述できるようになっている。

動詞の格構造記述は，係受け関係を処理する時点では，GRADEの辞書規則の形に展開されて使用される。したがって，その規則（部分文法，あるいは，部分文法ネットワーク）の適用モードを指定することにより，慣用句的な共起関係があらわれている場合には，その格構造記述の用法による解釈を，一般の共起関係で記述された用法よりも優先する等，柔軟な処理が可能である。また，使用頻度の多い和語動詞については，固定的なフォーマットの格構造記述では記述できないほどの多様な用法を持っているものが多く，これらの和語動詞については，その用法を分離するために，discrimination net に相当するGRADE部分文法ネットワークを動詞個別に定義し，それを辞書規則として呼び出す方式を採っている。この辞書規則は，従来埋込み型の手続としてLISP言語等で記述されていたものであるが，我々のシステムでは，この部分もまたGRADE規則という一様な記述形式で書かれることになる。

§8 構文解析とGRADE

ここで，本構文解析システムにおいてGRADEの特徴がどのように生かされているかについて述べておこう。本解析システムの特徴は，できるだけ「正解の可能性の高い解析結果」を唯一の解析結果として出力することである。このために，

(1) 実際のテキストを分析して得られるヒューリスティック規則を各段階で活用する
(2) 直接構文構造を組み上げる規則以外に処理範囲の限定だけを行う規則を使用する
(3) 構文構造の決定と同時に，意味解釈規則も構文解析の適切な段階で適宜行う（構文解析→意味解釈というように，2段階のフェーズ分けをしない）
(4) 単語個別の辞書規則を積極的に活用して

いる。そして，これらの性質の異る規則をすべてGRADEの木構造変換規則という一様な記述形式で書いていることである。一般に，このような種類の異なる規則を同一の枠組で取扱うことは，性質の異なる規則の混在による混乱を引きおこすことになるが，GRADEでは，規則の集合を部分文法として定義し，それを処理過程の適切な段階で起動（activate）する機能を活用することにより，この混乱を避けることができる。また，

(1) 多レベルの情報を処理の任意の段階で参照できる
(2) 性質の異なる規則を別々の部分文法の中に閉じ込め，各部分文法毎に，規則の性質に応じた適切な制御モードを指定できる
(3) 部分文法の適用順序をネットワークの形で定義できるので，辞書規則と一般規則のように，優先度の異なる規則をうまく取扱うことができる

という利点の方が大きいと考えられる。

また，通常の文脈自由型文法における書換え規則では，その規則によって1つの非終端記号に組み上げられるconstituentだけを参照するだけで規則適用の可否を判断するが，本解析システムで使われる各種のヒューリスティック規則（例えば，並列句・連用中止法のスコープ決定や連体形埋込み文のタイプ決定で使われる規則）は，

「周囲の環境をみて，最も妥当な解釈をどれかを判断し，最も妥当と思われる構造に組み上げてゆく」

という機能を果している。この種の規則は，複数個の文脈自由型文法での規則が組み上げるであろう構造を比較し，その中のどの構造がもっとも妥当であるかを予め判断する役割を果しており，文脈自由型規則の枠外であり，周囲環境（context）をチェックするというGRADEの機能が必要となる。

GRADEの各部分文法には，「多重解釈を出せ」というモードと「単一解釈で良い」というモードとが指定可能である。ほとんど誤解釈を出さない強いヒューリスティック規則の部分文法には「単一解釈」のモードを，逆に，誤りの多いヒューリスティック規則や文脈自由型文法的に動作させたい構造組み上げ規則からなる部分文法には「多重解釈」のモードを指定する，というように，複数個の構文解析結果を出力したい場合でも，文法設計者のキメの細い制御のもとにこれを行うことができる。また，ある部分文法ネットワークを多重解釈モードで実行させ，その結果をその部分文法ネットワークを呼び出した一段上のGRADE規則でフィルタをかける等の処理も可能になっているため，「妥当な解釈である可能性の高い」ものから順に出力してゆくことができる。

現在，本解析システムは６人の文法記述者によって各部分が担当され，開発されているが，このような多人数による文法開発環境においても，ＧＲＡＤＥの部分文法の考え方が有効であることが確認されている。

§9 おわりに

本解析システムは，現在，主要部分の設計を終了し，各部分のＧＲＡＤＥ規則化とデバッグを行っている。７月中に各部分のデバッグと全体的な統合を完了し，８月から形態素解析と結合して，解析実験を開始する予定である。

文献

(1) 長尾・辻井・田中・石川：「科学技術論文における並列句とその解析」，情報処理学会自然言語処理 36-4，１９８３．３．

(2) 長尾：「科技庁機械翻訳プロジェクトの概要」，情報処理学会自然言語処理研究会資料，１９８３．７．

(3) 坂本：「形態素解析」，同上

(4) 坂本：「格構造を中心とした用言と付属語辞書」，同上

(5) 中井：「語の収集と体言を中心とする辞書について」，同上

(6) 中村：「文法記述用ソフトウェア ＧＲＡＤＥ」，同上

(7) 奥津：「生成日本文法論」，大修館書店，１９７４

(8) 草薙：「日本語文解析におけるテンス・アスペクトの問題」，情報処理学会，自然言語処理研究会 34-12，１９８２．１２

(9) 辻井：「訳語選択について」，情報処理学会，自然言語処理技術シンポジウム予稿集，１９８３．６．