自動構文解析の構想

自動構文解析の構想

中野　洋

0．目的

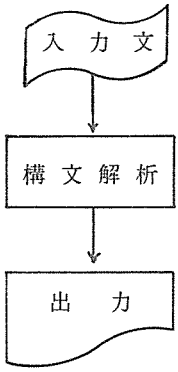
構文解析は言語情報処理の基礎的な作業である。機械翻訳や自動抄録，自然言語によるマン・マシン・コミュニケーションなど複雑な言語情報処理には必ずそれぞれの目的に応じた構文解析が必要であろう。しかし，ここでは特に，～のためのと掲げるような具体的な目的はない。しいてあげるなら，現在研究室で進行中の「漱石・鷗外の用語の研究」に文型索引を加えることだろうが，これもまだ具体化はしていない。

そういう具体的な目的とともに，筆者にとっては，人間の言語行動（特に文・文章の理解について）を機械にシュミレートしてみたいという欲求がある。学習機能や意味の処理を加えたのは，単に自動構文解析を可能にするためだけではなく，そういう欲求があってできたのである。

1．文法をつくるプログラム（プログラミング・プログラム）の考え

これまでの構文解析システムでは，まず文法を用意し，それにつかう辞書を用意し，それらによって処理するという方法であった。（図1）

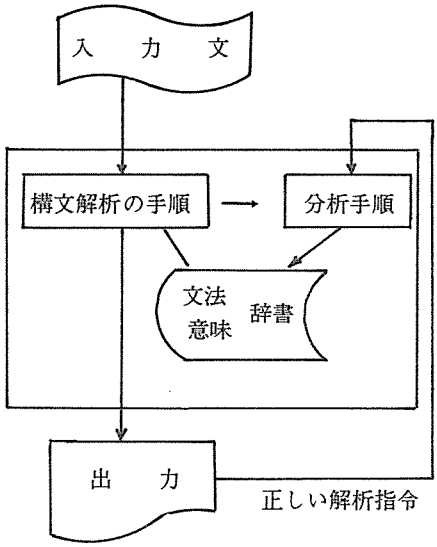
図1　これまでの構文解析



本論では，機械が入力文を分析し，その分析結果として，文法辞書や意味辞書をつくり，それにより構文解析を行なう。出力がまちがっている場合（この出力を使う人間が判断する），正しい解析指令を機械に対して出す。それにより，正しい出力を作り，文法辞書や意味辞書を正しくする。このシステムであらかじめ用意されなければならないのは，できる

だけ簡単にした分析手順と構文解析の手順である。つまりこのシステムは文法を作るプログラムと言えよう。（図２）

図２　ここでの構文解析システム



この方法は次のような考えによっている。

入力文が正しい意味を持った文章なら，語が無秩序に並んでいるのではなく何らかの規則によって並んでいるはずである。その規則は文法規則であり，語や文があらわそうとする表現対象がもっている体系（たとえば，地球上では，机の上に本がのり，机の下には本はのらない。―意味）からくる規則である。入力文を機械的に分析することによって，それらの規則を作り出し，作り出された規則によって構文解析をする。（小論文「言語情報処理における意味の把握の一方策」〈計量国語学53号〉参照）

２．最初の情報

ここでは，前述の小論での方法（二語文からはじめる。二語文は無条件で関係が決定する。その結果を辞書に貯めこんで，次の文にうつる。）はあまりに手間がかかるのでとらず，一番最初のデータにだけある程度の情報を入れておき，それを分析して辞書をつくる。次からは情報を入れないなまのデータを処理するという方法をとる。

ここでいう最初の情報とは語と語の関係である。

修飾の関係……いわゆるかかり・うけ

並立の関係

独立の関係……接続詞や感動詞の用法。他の特定の語とは関係をもたず，文や句・語の関係を示したり，話し手・書き手の態度をあらわす語。

〔例〕　しかし，彼と彼女は行った。

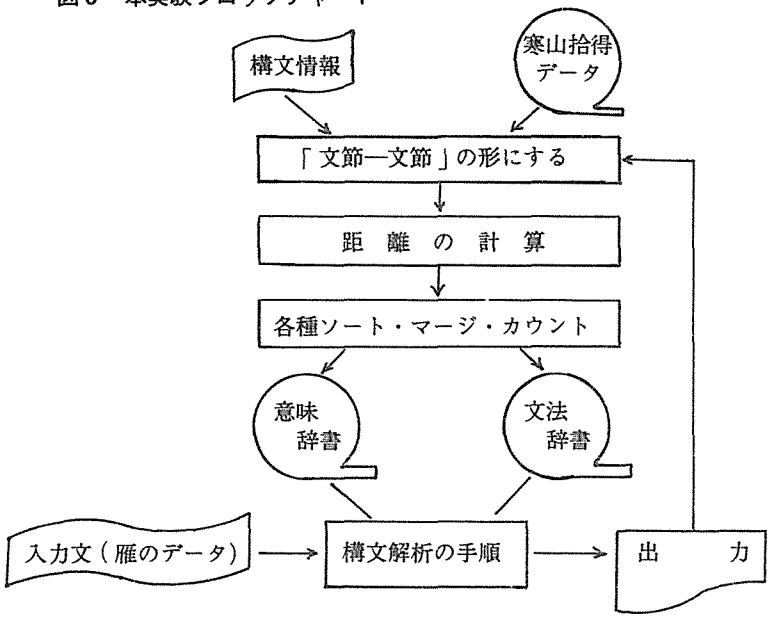
しかし……独立，彼ー彼女……並立，彼女ー行く……修飾

3．分析手順

ここでの分析は今のところ次の三点である。

1）ある語がどういう語にかかっているか，ある語がどういう語をうけているかを調べる。頻度数を計算する。……結果を意味辞書（うけ集合とかかり集合）とする。

図3　本実験ブロックチャート

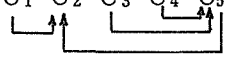


2）ある附属語（附属語がない場合は自立語自身）がどういう品詞にかかっているかを調べる。頻度数を計算する。…結果を文法辞書とする。

3）ある語が他の語と関係をもつ時，その距離はどれだけか。……結果を文

法辞書に加える。ここでいう距離とは，語のへだたりの数をいう。又，＋－で方向を示す。

〔例〕 $C_1\ C_2\ C_3\ C_4\ C_5$ の構造を持っている時，



$C_1 \rightarrow C_2$ 距離＋1

$C_3 \rightarrow C_5$ 〃 ＋2

$C_4 \rightarrow C_5$ 〃 ＋1

$C_5 \rightarrow C_2$ 〃 －3

第一資料研究室で作成ずみの，鷗外の作品「寒山拾得」（総のべ語数4021，総文節数2141）を最初のデータとして，辞書を作った。

本実験のブロックチャートを図3に示す。

4．辞書の内容

1）意味辞書

ある語をうける語の集合，ある語にかかる語の集合で，その語の抽象的意味が記述されていると考える。

1－1） かかり語見出し・うけ語集合（表1・かかり語見出し参照）

かかりの語（または並立の前に立つ語，または独立の語）が見出しで，それをうける（または並立にたてる）語が集めてある。見出しが文末の文節に使われた場合は空集合である。表1・かかり語見出しの「寒山」，附属語「う」「か」はこれである。

表を見て，次のようなことがわかる。「寒山」「拾得」は並立で用いられること。「と」を見ると「申す・言う・仰っしゃる・見える」があり，このうち「申す」は両見出しの中に見える。同じまたは同じ意味の語が並んでおり，「寒山」と「拾得」が同じような意味の語であることがわかる。名詞はいろいろな助詞を伴っていろいろな格に立てるが，「頭痛」はこのうち，ここでは「を・が・に・の・は」の格に立ち，表のような語を修飾できることがわかる。この表を用いれば，「頭痛が話す」のような文法的には正しいが意味的には正しくない文を作ることはない。

1－2）　うけ語見出し・かかり語集合（表1・うけ語見出し参照）

うけの語（または並立の最後に立つ語）が見出しで，それにかかる（または並立にたてる）語が集めてある。うけに立たない語（たとえば，連体詞・副詞など）はこの辞書の見出しとならない。

表を見て，次のようなことがわかる。「いれ」という語は「を」格と「に」格を取り，「みず・はち」「はち・つつ」という意味の名詞と修飾の関係にた

表1　意味辞書

① かかり語
② 頻度数
③ かかりについた附属語，またはかかり自身
④ 頻度数
⑤ 横文情報…修・修飾，並・並立，独・独立
⑥ うけ語
⑦ 度数

かかり語見出し

①(②)	③(④)	⑤	⑥　(⑦)…
寒山(10)	も (1)	修	来(1)
	う (1)		
	か (1)		
	寒山(2)	並	拾得(2)
	が (1)	修	言っ(1)
	と (3)	修	申す(2)仰っしゃっ(1)
	は (1)	修	まいる(1)
拾得(8)	は (1)	修	居ら(1)
	が (2)	修	普賢(1)洗ひ(1)
	拾得(1)	独	
	(1)		
	と (3)	修	申す(1)言ふ(1)見える(1)
頭痛	を (1)	修	取り(1)
	が (3)	修	する(2)おこっ(1)
	に (1)	修	悩ん(1)
	の (1)	修	ため(1)
	は (1)	修	あっ(1)

うけ語見出し

⑥(⑦)	③(④)	⑤	①　(②)…
いれ(4)	を (2)	修	みず(1)さい(1)
	に (2)	修	はち(1)つつ(1)
飲ん(4)	を (2)	修	水(1)薬(1)
	て (1)	修	かぶっ(1)
	で (1)	修	長安(1)
頭痛(4)	の (2)	修	レウマチス(1)ほど(1)
	その(1)	修	その(1)
	た (1)	修	い(1)

てる。「飲ん」は「を」格は取るが，「に」格は取らない（この表では）ことがわかる。したがって，この表で「水に飲んだ」という文法的に正しくない文を作ることはない。

〔例〕「頭痛」（「寒山拾得」のデータから）

あいにく，こらへられぬほどの頭痛がおこった。

実際，閭はこれまで頭痛がする，頭痛がすると気にして居て，どうしてもなおらせずに居た頭痛を，坊主の水に気を取られて，取り逃してしまったのである。

それに頭痛に悩んでおいでなさると申すことでございます。

単純なレウマチス性の頭痛ではあったが，……

その頭痛のために出立の日をのばさうかと思っていますが，……

「入れ」

汲みたての水を鉢に入れて来いと命じた。

残っている飯や菜を竹の筒に入れて取っておきますと，……

2）文法辞書

文節を構成する語の中で最後の語がその文法的性格を決定していると考え*，それがどういう品詞（文節中の最後の自立語の品詞がその文節の品詞であるとする）にかかるかを記した辞書。かつ，そのかかりの距離が記してある。（表2参照）

表2から次のようなことがわかる。「が」「は」はそれぞれ③の合計，106回，118回あらわれ，それぞれ名詞や動詞や形容詞などにかかった。（空白はそれが文末にあらわれたことを示す）「が」と「は」を比べると，「が」の方が「は」より動詞にかかる率が高いこと，動詞にかかる場合，「は」

表2　文法辞書

① かかる語についた附属語またはかかる語自身
② うける語の品詞
③ 頻度数
④ かかりの距離の平均値

①	②	③	④
が	副詞	1	1
	形容詞	1	1
		2	0
	固有名詞	3	1
	名詞	8	3
	動詞	91	2
は	固有名詞	1	1
	代名詞	1	1
		3	0
	形容詞	4	1
	名詞	19	3
	動詞	90	4
一番	形容詞	1	1
元来	動詞	1	13
いる	固有名詞	2	1
	名詞	10	1
		10	0

* 助詞連続で，格助詞ー副助詞の時は前の格助詞が働くなど，例外が多いが，ここではシステムを作ることが目的で，細部については後に検討する。

が「が」より遠くにかかることがわかる。副詞の中にも，「一番」のように近くの語にかかるもの，「元来」のように遠くにかかるものがあること，「いる」は名詞にかかるか，文末に用いられるか（つまり，連体形か終止形か）である。

3）辞書の量

この方法では辞書が厖大なものにならないかという疑問については，次のような見通しがある。文法辞書については，ふえる箇所は動詞，形容詞，副詞，連体詞，感動詞，接続詞など自立語単独で用いられる語である。新聞語彙調査一紙一年分，短単位のべ約百万語についての統計では，これらの語の総計は8837～6282語であり，これ以上はあまりふえないであろうと思われる。意味辞書については，すべて見出しに立つ（同じ一紙一年分の新聞で，固有名詞を除く自立語の総計は29822語である。）が，その内容は幾何級数的に増えることはない。ある語がすべての他の語と意味的に関係をもつということはなく，予想するより，非常に限られた範囲の語としか関係をもたないと思われる。又，技術的には，よく使われる語（頻度数の高い語）はよく用いる辞書に，あまり使われない語（頻度数の低い語）はあまり使われない辞書にというように，使用率で区分けして分けて納めること，特定の分野の文章を処理する場合には，その分野用の辞書を用意すること（辞書に層別の指標をつけておくなどする）などで実際に使う量をへらすことができる。

4）辞書の誤り

辞書の中には誤りがある。これは，入力時の構文情報が誤っているせいである。入力時の構文情報が完全に正しいものでなければならないのなら，この構文情報をつくるのは専門家でなければならない。我々が学習する時，誤りも習うはずであり，誤りを習っても正しくことばを使えるのは，その誤りをいつか指摘されて直すか，または，忘れるためである。このシステムでは，この二つの方法，誤りを直す，忘れるという機能をつけた。誤りを直すのは，出力に誤りが出た時で，正しい解析指令による。忘れる機能は，頻度数による。頻度数をつけることにより，正しい用法は頻度数が高くなり，したがってよく使われ，よく使われる辞書に納められ，まちがった用法は頻度数が低くなり，したがってあまり使われない辞書に納められる。よく使われる辞書が多くなればなるほ

ど，相対的にあやまった用法は使われないということで忘れるという機能がつく。したがって，入力時の構文情報をつけるのは専門家でなくてもよい。辞書の中に誤りがあってもよい。（最初の時期に，出力の検討を怠ると，誤った辞書項目によって処理され，その頻度数が高くなるということが起る。）

5）辞書の拡充

この方法の最大の特徴は，使用する辞書の誤りを直し，自動的に項目をふやし，内容を充実させることである。

新しい文章を処理すればするほど，辞書の項目はふえ（限度はあるが），内容は充実する。全く新しい語が出て来た時，その語に附属語がついていれば，文法辞書によって処理され，分析手順によって辞書に登録される。その語に附属語がついていない時は，構文解析の手順5により処理され，辞書に登録される。すでに処理されたことがある語は，新しい用法が登録されるか，頻度数が加算され，どの用法が最もよく使われるかが正確になる。

同じような語彙を使った文章なら文法辞書が充実し，全く異なる語彙を使った文章なら意味辞書も拡充される。

辞書のふえ方はこのシステムの学習課程である。これ自身，興味ある研究対象になろうが，今は言及できない。

5．構文解析の手順

構文解析の手順はできるだけ簡単な方がよい。細かいことは辞書に従うという方針である。次の五つのステップで構成される。

0．かかる語より後の語をしらべる。

1．文法辞書をしらべる。可能性をすべて出す。

2．意味辞書をしらべる。可能性をすべて出す。

3．1と2どちらも満たす語が一つあれば，それにかける。二つ以上あれば距離による。それでも二つ以上あれば近い方にかける。

1を満たす語と2を満たす語が別の語であれば，1を満たす語をとる。

1を満たす語が二つ以上あれば，距離による。それでも二つ以上あれば頻度数の多い方をとる。それでも二つ以上あれば近い方にかける。

2を満たす語が二つ以上あれば，近い方にかける。

4．3で決まらない場合（辞書にない場合）は，その語はそのままにして，次の語にうつる。1へ。

5．1～4の処理が一文について終れば，かかりの矢線は交差しない，かかりは後の語（二語以上あればより近い語）にかけるという規則をあてはめて，決まらなかった語のかかりを決定し，間違いを直す。

この構文解析の手順によって，実際に構文解析をしてみよう。入力文は単位切りされ，品詞情報がついているものとする。この方法では辞書が完全であるかどうかが処理結果に大きな影響を与える。二つの辞書の状態によって，四つの場合が考えられる。文法辞書・意味辞書ともに完全な場合。「寒山拾得」のデータを再び処理する時や，「寒山拾得」の語彙を使った文——抄録など——を処理する時がこの場合である。文法辞書だけがよい場合。すべての名詞類を除く自立語が登録されていないと文法辞書は完全にはならない。ほとんどがこの場合に入る。意味辞書だけが完全な場合は考えられない。意味辞書に登録されていれば，文法辞書にも登録されているはずである。文法辞書も意味辞書も不完全な場合の構文解析は直後の語にかかるという文構造の結果だけしか出さない。

(1) 文法辞書も意味辞書も完全な場合

はちに　水を　入れる

まず，「はちに」を処理する。「に」がかかりえる語は文法辞書により，この場合，名詞（水）と動詞（入れる）である。又，「はち」がかかりえる語は意味辞書により「入れる」である。どちらも満足する「入れる」をとり，「はちに」は「入れる」にかける。同様にして，「水を」は「入れる」にかける。

(2) 文法辞書だけしか使えない場合

慰めるやうに　お玉の　顔を　見て　起ち上がる

まず，「慰めるやうに」を処理する。「やうに」は文法辞書により「見て」「起ち上がる」にかかりえる。「やうに」が動詞にかかる時の距離は1であるから，その±2を満足する「見て」にかける。「お玉の」は「の」の文法辞書により，「顔を」「見て」「起ち上がる」が選ばれる。文法辞書の頻度数の多

い方をとるという規則で名詞をとり，「顔を」にかける。「顔を」は文法辞書により，「見て」と「起ち上がる」が選ばれるが，近い方にかけるという規則で「見て」がとられる。「見て」は「起ち上がる」にかけられる。

この例では正しく解析されたが，いつもうまくいくとは限らない。というより，現段階では（「寒山拾得」の文章を分析しただけでは）ほとんどは一文に一箇所は誤りがあるという程度だとおもわれる。その理由は，「～について，～に関して，～に対して，～という，～かも知れない，～てしまう，～てある，～ている」など，二文節にわたっていつもあらわれる用法を切って処理していること，決まらない時はどんな場合でも前の語にかけていること（「は」などは後にかけた方がよい。）などが大きな原因であろう。

また，今回は，一語と一語の関係を前から決定していく方法を取ったが，後から決定していく方法や，文字連続*や語連続と構文の関係**を利用する方法，一語一語を決定するのではなくて，句を決定していく方法（一語一語の関係を決定し終えた後で，その結果を利用して句を決定することはできるが，これは一語一語の関係を句を決定しながら決定していく方法である）などが考えられる。どれか一つを取るというのではなく，長所を利用してゆきたい。今後の課題である。

6．おわりに

ここでは構文解析の方法を中心に述べたが，まだ，正しい解析指令の出し方，その効率，意味辞書の用法で見出し語の二次元的な使い方（ある見出し語内にはないが，同じ用法の他の見出し語内にはある場合，それを使う方法），これらの辞書を使っての文の作成，我々が普通言っている意味をこの方法でどの程度まで処理できるのかの研究，うまくいけば分類語彙表の自動作成，この構文

＊　ＬＤＰ10　斎藤秀紀「漢字かな混り文の文字列」参照

＊＊　本論集　田中論文「句のエントロピーにもとづく構文合成」，霜岡論文「電子計算機による代表構文作成の試み」参照

解析システムの具体的な利用法としての文型索引の作り方など，大小さまざまの，筆者にとってたいへん興味のある問題がふれられていない。次にはこれらの問題についても考えてゆきたい。

樺島忠夫氏，林四郎氏，高橋太郎氏や当研究所の第一資料研究室・言語計量調査室，第三資料研究室のメンバーその他の人々から，多くの貴重な助言をいただいた。まだまだ消化しきっていない点が多いがこれからの研究に取り入れてゆきたい。記して感謝の意を表する。

(1973年 8 月31日提出)