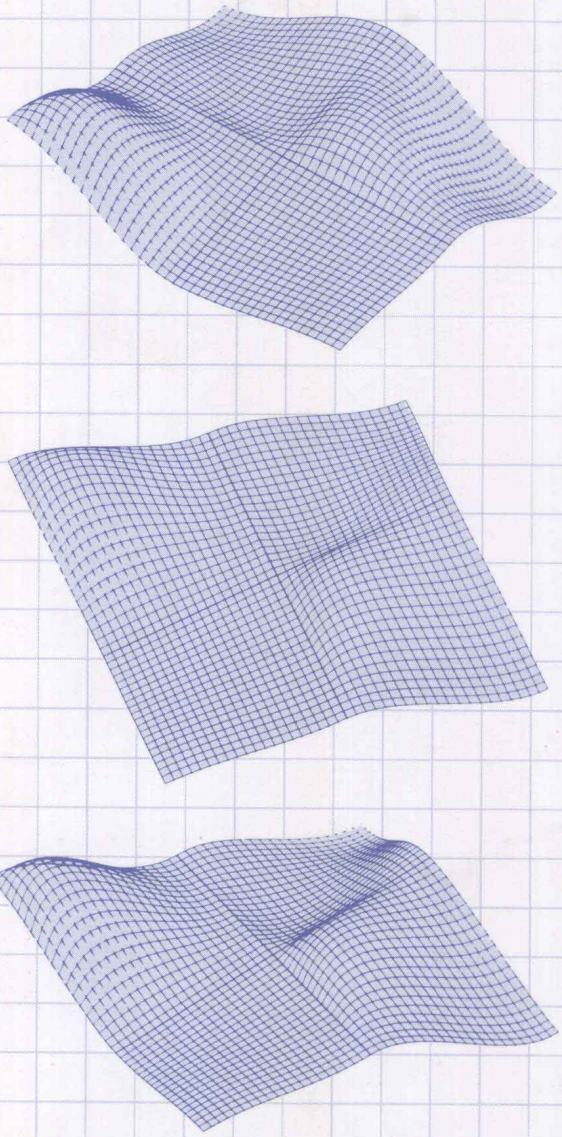
コンピュータ言語学入門

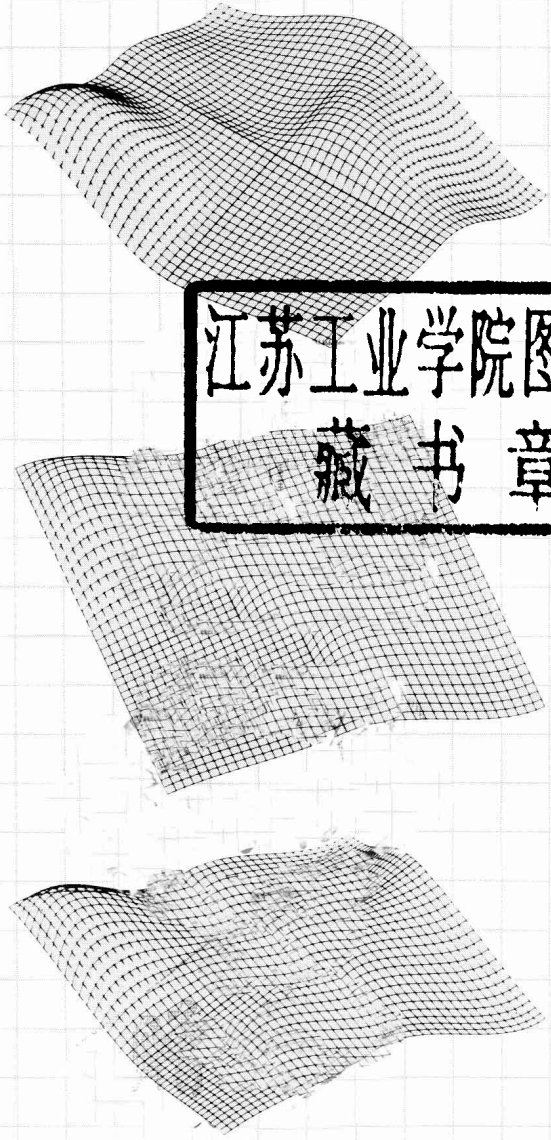
草薙 裕 著



大修館書店

コンピュータ言語学入門

草薙 裕 著



大修館書店

草薙　裕（くさなぎゆたか）

1936年台北に生まれる。1960年上智大学外国語学部英語学科卒業。1970年ジョージタウン大学大学院博士課程言語学研究科卒業，Ph. D.取得。ジョージタウン大学語学・言語学部講師，ハワイ大学文理学部助教授，同大学准教授を経て，現在筑波大学文芸・言語学系教授。著書：*Comprehension in Japanese with Emphasis on Aural Skills*, Vol.1,2 (CALM, U. of Hawaii), *Japanese Linguistics and Language Teaching*（編著，Dept. of EALA, U. of Hawaii）などの他に言語学に関する論文が多数ある。

コンピュータ言語学入門

1983年3月1日　初版発行
1988年5月20日　3版発行
定価1800円

検　印
省　略

著　者　草　薙　　裕
発行者　鈴　木　荘　夫

発行所　株式会社　大　修　館　書　店

101　東京都千代田区神田錦町3-24
電話東京(294)2221(大代表)振替　東京 9-40504

印刷／壮光舎　　製本／関山製本社

ISBN4-469-21109-5
Printed in Japan

序

今から二年半ぐらい前のこと，偶然本屋でマイクロ・コンピュータの月刊誌が目にとまった。頁をめくって行くうちに，驚きから大きな期待へと変わっていった。

それまでマイクロ・コンピュータと言えばマニアがハンダ付けをして作る，我々にとってはほとんど役に立たないシロモノだと思っていた。

ところが，マイクロ・コンピュータに大型コンピュータ用とほとんどかわらないキーボードがつき，プリンターも接続できるパーソナル・コンピュータは記憶容量が少し前の中型コンピュータより大きい。

しかも大学の研究費でも，自身の予算でも購入できる程度の値段ではないか。

早速，パーソナル・コンピュータで使えるプログラム言語（第一章参照）を調べた。そして市販されているパーソナル・コンピュータのいくつかは，言語の処理に必要な文字列処理が十分出来る機能を持っていることがわかった。

まもなく，私の研究室と自宅の机の上にパーソナル・コンピュータが備わった。それ以来ほとんど毎日コンピュータとの“対話”を行っている。

それまでの十数年間，勤めている大学や客員で招かれて行った大学のコンピュータ・センターの大型機との“付き合い”をして来たが，面倒な事務的手続き，センターへ出向く時間，待ち時間などの無駄，複雑な手続き言語などを考えると同僚の言語学者や学生達にコンピュータ利用を勧めるのが躊躇された。

ところが研究室や自宅のパーソナル・コンピュータは文字通り，パーソナルなもので，自分の好きな時にスイッチを押せばすぐ使える。プログラム言語もごく簡単であるにもかかわらず，文字処理も十分出来るし，拡張して性能を上げることも出来る。

まわりの同僚や学生達にパーソナル・コンピュータの効用を説いてまわっている。

私がパーソナル・コンピュータを使い始めてからも世の中がすさまじい勢いで変わって来た。大規模集積回路（第一章参照）を中心とした技術の進歩と市場拡大による大量生産の結果，性能が向上し，値段もどんどん下がって来た。

世はあげてマイクロ・コンピュータ時代，パーソナル・コンピュータ時代，オフィス・オートメーション時代だと業界やマスコミがぶち上げている。

ところが，一方では，雨後の筍のように出来たマイコン・スクールでは受講者1000人に対して，自分でプログラムが組めるようになるのは6人だけだという伝説や，時代に遅れまいと会社が導入したコンピュータの多くがホコリをかぶっているという話などを耳にする。

せっかく身近かになったコンピュータを活用出来ないのは，作られたブームに乗り遅れまいと，自分の仕事では何がコンピュータにさせられるかという認識や，何をさせようという目的がなく，ただ漫然とコンピュータの前に座るからではないかと思う。

言語学に関して言えば，言語の研究の根本的な問題，すなわち，言語とはどういうものか，目的は何か，分析や記述の過程での形式化などの問題から考えて行くべきだと思う。

そこで，言語の研究（言語学，国語学，英語学，フランス語学，その他，何々語学といわれるすべてに亘って）を志す学生諸君や言語に興味を持つ人々向けに現代の言語学の根本問題を考えながら，普通のクラスや入門書ではごく当然なこととして教えられたり，書かれたりするような言語の構造，言語の規則化，枝分かれ図などの概念を全くの基礎から，わかりやすく考えて行き，その延長として，コンピュータ利用へ結びつけて行くという私の希望がこの本となった。

言語の根本概念からコンピュータ利用へと行くには，当然，数学の基礎概念が橋渡しとなる。文科系の人の中には数学とかコンピュータは，はじめから自分には向かないというアレルギーを持っている人が多いようだ。ところが扱う材料である言語も，記号の体系の一つだし，また，それを処理する現代言語学の方法論も文科の中では一番理科系の発想に近いものである。

言語の好きな人は，数学と同じ様に，記号を操っているのだし，言語の研究に進んだ人は，やはり記号の体系の研究をしているわけで，アレルギ

ーがあるとしたら，直観的に理解しにくい数学の記号に抵抗を持っているのであろう。

本書で扱う数学は計算とか証明はいっさい含まれていない。むしろ抽象的な数学の記号と具体的な言語をどういうように結びつけるかといった論理の展開が主になっている。

こういう基礎的な事柄を理解することが，言語学各論のよりよい理解，個別言語学への応用などの手助けとなるのではないかとひそかに考えている。

そういう意味で，本書は，コンピュータのプログラムの具体例はのせてはあるが，プログラムの手引き書ではない。本書でコンピュータを言語学で用いる意味を理解し，実際に用いようという方は巻末の文献を参考に実際のプログラム入門書を勉強されることを勧めたい。

本書を書くにあたって，直接，間接に多くの御教示を受けたオランダのグローニンゲン大学文学部の D.G. ステュワート教授，京都大学工学部の長尾真教授，国際キリスト教大学教養学部の野崎昭弘教授，東京女子大学文理学部の水谷静夫教授をはじめ多くの方々に心よりお礼を述べたい。

次の各社から資料の提供を受けた。ここに感謝する。

エーペックス，富士通，日立家電販売，伊藤商事，高電社，三菱電機，日本電気，生活構造研究所，シャープ，信州精器，システムズフォーミュレート

また，日本電気パーソナルコンピュータ事業部の杉崎寅男氏，同土浦支店の阿波野洋氏，I/O ポート筑波の太郎良嘉親氏にいろいろお世話になった。

大修館書店の言語編集部の山本茂男，藤田佻一郎の両氏には殊の外お世話になった。特に藤田氏の助言がなかったら本書は実現しなかったであろう。ここで改めて御礼申し上げたい。

なお本書にのせたコンピュータのプログラムはすべて日本電気のパーソナル・コンピュータ PC 8801 および PC 8001 を用いたものである。

1982 年 12 月 25 日

著者

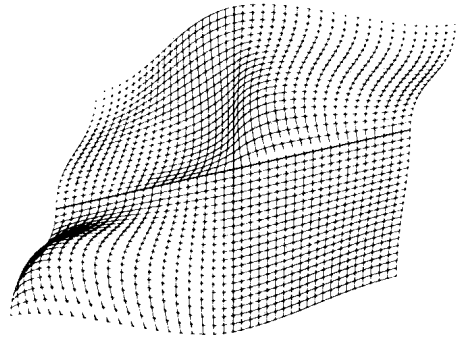
目　次

コンピュータグラフィック　**三井秀樹**（筑波大学芸術学系講師）
矢野研策（筑波大学情報学類学生）

装　　幀　　**鳥居　満**

BASIC 関係用語

コンピュータ言語学入門



第1章　プロローグ

1.　コンピュータ言語学

コンピュータ言語学とは一口に言えばコンピュータを用いて言語の研究を行なうことと言えよう。

ここで言う言語とは日本語とか英語とかスペイン語とかタイ語といった我々人間が日常生活に用いる言語のことで，これを特にコンピュータの世界では自然言語ということがある。これは，コンピュータとのやりとりを行なう一連の命令のことも「言語」と呼ぶからである。

コンピュータを用いる自然言語の研究は目的に従って二つに分けられる。

一つは，言語の構造や言語使用を解明するもので，コンピュータをその理論を確かめるために用いるものである。

もう一つは，言語の使用をより効率よくするためにコンピュータが用いられるようにする目的で言語について研究することである。

この両者の間には，もちろん，深いかかわりがあり，互に影響し合うこともあるが，前者の目的はあくまでも言語とは何か，言語はどのように用いられるかなどを解き明かすことが目的であるのに対し，後者は，我々人

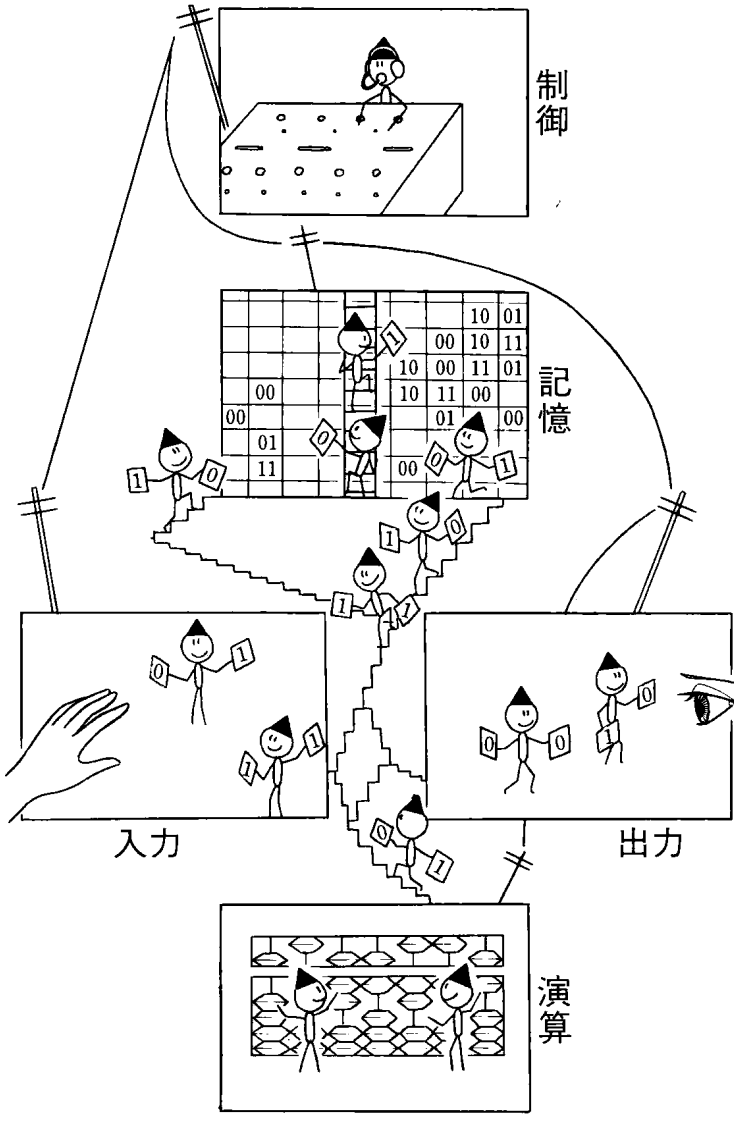


図1.1 コンピュータの機能

間がことばを使う時にコンピュータに手助けをさせようという，いわば開発のための研究が目的であるので，コンピュータの特質に大きく影響される。

本書では，言語の研究のためのコンピュータ利用を中心に考えていく。

最近の言語学は構造の形式化が進んでいる。これは，音であれ文法であれ意味であれ，構造を記述する際，普通のことばを用いれば説明があいまいになったり不完全になることが多い。だから厳密に定義された記号を用いて明示的な規則で構造を記述するという方法が用いられるのである。

言語の規則が形式化されれば，後で述べるようにコンピュータへの指示が作りやすい。またコンピュータは人間の作った指示通りに動くので，形式的に記述した規則をコンピュータで検定してみると，その規則が適切かどうかの判定がたやすく行なえる。

本書は言語の形式化の根本的な問題をとりあげ，それを出来るだけわかりやすく説明しながら，それと関連したコンピュータへの命令の手続きを考えて行く。

2. コンピュータの構造

まず最初にコンピュータを概観しておこう。

コンピュータの機能は，それが大型のものであれ，超小型のパーソナル・コンピュータであれ，（1）入力，（2）記憶，（3）演算，（4）出力，（5）制御，の五つということができる。

コンピュータは数でも文でも，入れる，覚える，変える，そして，出す。また，それらの動作を動かす，と思えばいいだろう。その仕組みは図1.1に示してある。人間もたとえば文を耳から入れ，一時，覚え，その形を変える，そして口から声で出す，といったことをやる。それらの動作を動かすのは神経だ。数や文を変えることでは人間もコンピュータもあまり変わらない。

コンピュータの入力と出力はいわば人間とコンピュータのやりとりのことで，入力は人間がコンピュータに意志を伝えること，出力はコンピュータが人間にその操作の結果を伝えることだと思えばよい。

記憶は入力されたものを一時的にたくわえること，演算は，それをいろいろの形に変えること，そして制御は，入力，出力，記憶，演算などに用いられる装置を動かすことである。

コンピュータがこれらの機能を発揮するためには一つひとつの操作の手順に対する指示がなければならない。これをプログラムという。

コンピュータはプログラムが記憶装置に入れられており，それによって自動的に動作する。これをプログラム内蔵方式という。

コンピュータは電気の機械であることは言うまでもない。ではコンピュータは数や文を内部でどの様に処理しているのであろうか。

それを述べる前に「情報」という言葉を考えておこう。「情報」は日常の言葉では，ある物事に関する知らせや知識といった意味で用いられているが，ここでは，コンピュータに処理されるいろいろの記号の意味付けというように使うことにする。

コンピュータで処理を施される数や文はすべて情報というわけである。

コンピュータには数や文が情報（これをデータとも言う）として入り，プログラムにより，その数や文がいろいろと形を変え，違った形の情報として出てくることになる。

次に，人間が意味付けした記号，すなわち情報がコンピュータの内部でどのように処理されるかを考えよう。

コンピュータは電気の機械だから情報も電気や磁気に置きかえられる。その秘密は電気のオン・オフ，あるいは，磁気の＋・－にある。

すべての情報を細かく分け，それを白か黒かにするわけだ。これは人間の感覚で言えば，1と0のいわゆる二進法なのである。

我々日本人（何も日本人に限ったことではなく，世界の人々の多くもそうだが）は0から数えて9まで行き，次の数で桁上がりにして，10になる

十進法	二進法	十六進法
0	0	0
1	1	1
2	10	2
3	11	3
4	100	4
5	101	5
6	110	6
7	111	7
8	1000	8
9	1001	9
10	1010	A
11	1011	B
12	1100	C
13	1101	D
14	1110	E
15	1111	F
16	10000	10

表1.1

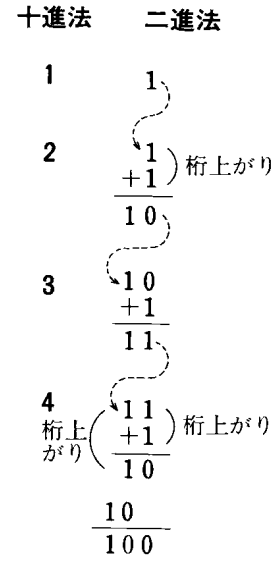


図1.2

十進法に慣れている。

したがって二進法になると，慣れるまでやや面食らうが，この数え方は数字が0と1としかなく，1の次，すなわち十進法の2になると，桁上がりして10とする。そして十進法の3は，一桁目が0の次だから，1で11，次の十進法4は，一桁目が桁上がりし，その桁上がりが二桁目をさらに桁上がりさせて，結局100になる。

すなわち，下の方の桁から十進法の1，2，4，8……というようになる。この1，2，4，8……は，いいかえれば2^0，2^1，2^2，2^3，……となっているわけである。

二進法では同じ桁で0と0を加えれば0，1と0を加えれば1，1と1

を加えればその桁は0で一桁上が1になる（図1.2）

コンピュータの内部では情報も命令もこのような1と0のみの数字の集まりのような形で記憶されたり，理解されたりしていると考えればよい。この1か0の一桁が情報の最小単位というわけだ。この最小単位をビット（bit）という。これはbinary digit（二進法の桁）の略である。

さらに，この1と0の組み合わせのビットが8とか16とかを単位として用いる。

我々人間にとっては，1や0の集まりを覚えたり使ったりするのは非常に不便であり困難なので，もしそのようなことが必要な場合は普通，十進法の0から15までを一桁として表わす十六進法にする場合が多い。十六進法は1から始めて9までは十進法と同じだが，十進法の15までは桁上がりしない。したがって十進法の10から15までは十進法では二桁の数だが，十六進法では桁上がりしないので，10……15のように使うわけにはいかず十進法の10をアルファベットのA，11をB，というようにあて，以下，C，D，Eとして行き，十進法の15をFで書き表わすわけである。この対応は表1.1に示してある。

ところで，二進法では一桁目で十進法の1まで，二桁目で3まで，三桁目で7まで，四桁目で15まで表わす。すなわち二進法四桁が十進法の0から15までを表わすわけで，これは十六進法の一桁と一致する。だから機械の処理と合う二進法をさらに人間に分かりやすく，あるいは覚えやすくするのに十六進法が都合がいいわけで，人間が直接コンピュータに情報を入れる場合も二進法ではなく十六進法で入れるのが普通である。

ただし，後で述べるように，二進法や十六進法を用いてコンピュータに仕事をさせるのはごく特殊な場合で，我々の目的である自然言語のためのコンピュータ利用ではデータの型との関連程度に十六進法が必要になるくらいである。

3. コンピュータの装置

コンピュータにいろいろの仕事をさせるためには，その手順を示したプログラムやデータを，コンピュータの中心であり，制御，演算，記憶をつかさどる中央処理装置（CPU—Central Processing Unit）に入れ，その仕事の結果を人間が理解出来る形で出させなければならない。

入力と出力の装置が人間とコンピュータの情報の受け渡しをする。

大型のコンピュータの処理の速さは人間の動作に比べればとてつもなく早く，たとえば，タイプライターのようなキーボードに人間が情報を打ち込むとすれば，その間，コンピュータは処理をしないで遊んでいることになる。そこで最初は人間が情報をコンピュータに理解できる形にする操作とコンピュータの処理を分離した。

人間は前もって一定のカードに情報に対応した穴をあけておく。そしてその穴のあいたカードをたくさん作っておき，いっぺんにカード読取装置で読ませるという方法だ。

カードに穴をあけるためには，タイプライターに似たキーボードを備えたカードせん孔装置を用いる。また紙テープせん孔装置もある。

最近はその速度の違いを利用したタイムシェアリング・システム（TSS—Time Sharing System）が普及している。これは大型コンピュータを複数の利用者が共同して用いるものであり，通常，利用者に都合のいい場所，したがって，コンピュータ・センターばかりでなく，利用者の仕事場の近くに端末機を置き，キーボードから直接情報をコンピュータに入力する。

コンピュータは時間のかかる入力や出力を行っている間，遊んでいないで，多数の利用者の仕事をどんどん処理していく。コンピュータの処理速度が桁違いに速いから，利用者の方はあたかも自分一人でコンピュータを利用しているように思える，というのがTSSの利用形態である。

最近の技術の進歩で，所定の位置に鉛筆で印をつけたり，字を書いたり