〔翻　訳〕

「コンピュータ言語学の課題」[訳注 1)]

ローラント・ハウサー
山 田 善 久（訳）

20 世紀の 40 年代に最初の電算装置が開発されて以来，数値的情報論と非数値的情報論が区別されている〔傍点箇所は原文イタリック体。以下同じ〕。数値的情報論は数の計算にたずさわり，物理学，化学，経済学，社会学などにおいて，爆発的な知識の拡大に導いた。一方，銀行業，航空交通，在庫管理などの多くの実用的な領域においても，今日もはや数値的応用抜きでは考えられない。コンピュータとそのソフトウェアなしでは，これらの領域の作用能力は破綻するだろう。

他方，非数値的情報論は，知覚や認知の現象にたずさわる。希望に満ちた開始にもかかわらず，非数値的情報論の理論的および実際的発展は，数値的情報論のそれのはるか背後にとどまっている。しかしながら，近年，非数値的情報論は，**認知科学**および**人工知能**として再び強い関心を見出している。英語で言うところの “cognitive science” および “artificial intelligence” は，情報科学，心理学，言語学，哲学，記号論理学を包含しつつ，人間の情報処理を研究する。

1. 方法と応用

コンピュータ言語学においては，理論的仮説を組織的にコンピュータ・プログラ

ムとして実現するという点で，理論言語学，心理学，哲学および記号論理学の非常に異なった諸方法が，一つの共通項にまとめられる。このコンピュータ上での理論的仮説の展開や検証は，従来の言語学，心理学，哲学および記号論理学とは明確に区別される新しい統一的な方法論と理論構築の可能性を提示する。

コンピュータ言語学の発展は，理論面では，新しい方法論（コンピュータの組織的導入による形式文法体系の検証）によって，実用面では，有用な自動言語処理が格別に必要とされていることによって推進される。現代言語学の日々の科学研究において，コンピュータの組織的導入は，次のような効用をもたらす。

1.1 プログラミングの方法論的効用

▽ プログラミング，例えば解析パーザーとしての文法のプログラミングには，取り扱われるべき現象を，以前ふつうに行われていた以上に，ずっと詳細に分析することが必要となる。

▽ プログラミング向きの文法形式論といった従来とは異なる適性が，競合する諸理論の張り合いの中で，新しい重要なファクターになる。

▽ 効果的に実現可能な（および実現された）自動言語分析の文法形式論は，実際面での応用を持ち，この応用は，まったく新しい独特のダイナミズムを精神科学の経験的知識にもたらす。

以下の実際面での応用の諸分野においては，コンピュータ言語学の方法が，ますます大規模に導入されつつある。

1.2 コンピュータ言語学の実際的課題

▽ テキスト･データベースのインデックス化と呼び出し

テキスト･データベースは，テキストを電子の形式で保存する。例えば日刊紙の各年度分，医学雑誌の公表論文，あるいは 1960 年以降のアメリカ合衆国での全裁判所判例など。こうしたデータベースの利用者は，自分の特殊な問題設定にとって重要なすべてのドキュメントやテキストの該当箇所を見つけることができるに違いない。

▽　自動テキスト生産

エンジンとかポンプとかテレビなどのような絶えず新しい製品を発売する大企業は，このためにくりかえし新しい製品説明書や保守マニュアルを作成しなければならない。同じことは，非常に多くの文書のやりとりをする弁護士，税理事務所，人事課などにもあてはまる。この場合，各手紙は内容の違いを明確に示す箇所でのみ区別される。テキスト生産の方法は，単純なひな型によるものから，言語学の知見に基づく高度に柔軟性のある双方向のシステムにまで及ぶ。

▽　自動テキスト点検

この分野でも，応用は，語形リストに基づく単純な正書法のチェックから，造語の領域でのさまざまな言語現象を組織的に取り扱う形態素システム，さらには語順や一致（Kongruenz）などの誤りを見つけることができる文章チェッカーにまで及ぶ。

▽　自動内容分析

活字になった情報は，この地球上で，10年ごとに，2倍になると言われる。法律，経済などの学術専門分野においても，絶えず新しい状況にいるためには，研究者の一生ではもはやまったく間に合わないほど，重要文献が氾濫している。短い要約付きの信頼のおける自動内容分析が，ここでは，きわめて有益であろう。自動内容分析は，テキスト・データベースから最適な呼び出しを行うために必要であるような概念ベースによるインデックス化の前提でもある。また，実効のある機械翻訳の前提である（以下参照）。

▽　機械翻訳

機械翻訳は，非数値的情報論の始まりの頃には，主要応用分野の一つであった。1955年から65年の10年間に，この分野で，集中的に研究がなされ，世間から大いに注目された。しかしこの期待は実現しなかった。商業的成功への期待は打ち砕かれた。

その間に関心は再び著しく高まった。HUTCHINS 1986は機械翻訳が引き続き研究されることに対し，次のような理由をあげている。

—— 学者，技術者，技師，経営者およびその他多くのビジネスマンは，日々，自分がマスターしていない言語で，たくさんの手紙や文書を読んだり，書いたりしなければならない。こうした絶えず増えている資料の山を処理するためには，翻訳者の数がまったく足りない。

—— 多くの研究者は，理想主義から機械翻訳の開発にたずさわっている。彼らは，機械翻訳の開発が言語の壁を克服し，工学，農学，医学の知識の発展途上国への普及をうながすことから，国際協力や平和を促進しようとしている。

—— 一方，機械翻訳は軍事分野での応用をもくろむ機関によっても促進されている。例えば，敵の文書の迅速な翻訳など。

—— 純粋な研究の課題としては，機械翻訳は困難な課題である。これの解決を言語学研究のテストとみなすことで，研究にたずさわる学者もいる。

—— 最後に，性能の高い自動翻訳システムは，広範な応用を持った価値あるソフトウェア製品であり，これによって多額の金を稼ぐことができるため，開発される。

現在12の参加国の中で九つの異なる言語が話されているヨーロッパ共同体においてまさに，自動または半自動の翻訳システムの持つ潜在的効用は計り知れない。

▽　自動化された授業

多くの時間が，いわゆるドリル学習のために費やされる数多くの授業科目がある。例えば語学の授業における，多かれ少なかれ機械的な規則的および不規則的変化表の習得。これらはコンピュータで，少なくとも同じ程度には，実施することができる。これによって，教師には，他の活動，例えば会話のためにより多くの時間が残される。最近の研究においては，集中的に，授業のさまざまな局面での自動化の問題が取り扱われている。例えば翻訳練習の際の自動誤り分析など。

自動化された授業システムは，自動的に生徒とコンピュータのやりとりを記録することができるという補足的な利点を持つ。生徒がどこで一番多く間違いをおかすか，どこで一番多くの時間がかかっているかについて知ることにより，自動化された授業の人間工学的側面の改善のための価値ある発見的手がかりが得られる。これは旧式の「電子テキストブック」から，メディア対応の独自の教育法を持つ新式の教育ソフトへの発展に道を開いた。

▽　対話システムと自動案内

コンピュータとのインターアクションの際の本質的な難点は，このインターアクションが費用のかかるものである（例えば自分で開発したプログラム），あるいは柔軟でない（例えばメニュー操作のインターアクション）という事実である。それゆえ，実際のあらゆる人間と機械のインターアクションの際に投入可能な，しっかりした自然言語システムの開発に大きな関心が寄せられている。

コンピュータ言語学の可能な応用分野の数は，これで決して完結するわけではない。むしろそれは，ごく一般的に，（今日そして未来において）人間がコンピュータと関わる全領域をカバーする。これらすべての応用分野において，言語学の知識が，少なくとも潜在的には，自動言語処理の最適化に役立つのである。逆にコンピュータは言語学的分析および理論形成の補助手段として，ますます大きな役割を演じることになる。

2. 電子メディアへの転換

自然言語をコンピュータ上で自動的に分析したり，処理したりすることができるためには，それを電子的に表示された文字列として記憶する必要がある。こうした形式でコンピュータ上に記憶されたテキストは，オンライン・テキストとも呼ばれる。

しかしながら自然言語のテキストや語句は，たいてい，さまざまな非電子メディアのかたちで存在する。話しことばの音声記号，書きことばや印刷された言語の文字，あるいは聾唖者の言語の身振りなど。音声記号と身振りは，（テープ録音やビデオ録画はとりあえず度外視して）通常，束の間の寿命しかないのに対して，文字は，伝統的に紙や羊皮紙や石に定着され，その断然長い耐久性の点においてすぐれている。

文字言語の伝統的な記録に対して，電子メディアにおいては，磁気テープ，ディスクあるいはCD-ROMがデータの担い手として使われる。非電子的に記録された言語を電子メディアへ転換するには，費用がかかるが，いろんな方法で行うことができる。

一つの可能性は，話しことばないしは書きことばを（例えば秘書を通じて）コン

ピュータにタイプ入力することである。これは今日なお広く普及している方法で，例えば，オフィスでの口述用テープレコーダによるタイピング，心理学におけるテープ録音のトランスクリプション，従来は印刷された形でしか存在しなかった書物の入力などがある。

それに加えて今日では，テクノロジーに基づく方法が現れている。（紙の上に）印刷された言語を電子メディアに自動的に移送することは，光学的パターン認識の分野に入り，いわゆるスキャナーを使って行われる。この機械は，写真でも行えるようなページのコピーのみならず，個々の文字を行ごとに走査して，それを記憶されているパターンと比較する。このようにして印刷イメージは，（いわゆるビットマップとして）コンピュータにコピーされるだけでなく，文字単位で**認識**される。これが光学的文字認識ないしはOCRのプロセスである。

ところで印刷イメージは，本ごとに非常に異なっている可能性がある。さらに，表題，脚注，図の解説あるいは表といった部分では，文字の大きさや書式が異なってくる。最近のスキャナーは，ある特定の文字について，それが例えば ‘a’ なのか ‘d’ なのかをプログラムに入力することにより，ユーザが誤認識を訂正することができる予備学習段階を用いて，これを克服している。

さらに高機能スキャナーは大きな辞書を使用し，それによって，疑わしい箇所で，二つの可能性のうちどちらが有意の語形なのかを決定する。このようにして，使用されている文字タイプや文字イメージの品質に依存せずに，99% までの認識率を達成することができる。その際，装置は，1ページに対して，50秒から数分を要する[1)]。

人間による本のページのタイピングと比較して，現代のスキャナーのスピードは競争に耐え得るものである。とくに機械は疲れを知らず，スキャナーの操作は，習熟していない人手で行えることを考えた場合には。しかしもっとも重要なファクターは，誤りがなくなることである。しかしこの場合でも，重要な文書の際には，あとで校正読みがなされなければならないために，両方の転換形態は必要である。

スキャナーとそのOCRソフトのパフォーマンスは80年代以来，ものすごく改善されてきた。コンピュータ分野で特徴的な価格の下落を同時にともないながら。そのため1991年以来，オフィスでのスキャナーの普及が著しく高まっていることが観察できる。

一方，話しことばを電子メディアに転換することは，ずっと難しい。印刷イメー

ジが，わりあいつぶのそろった文字を持つ明瞭に区切られた単語を保持しているのに対して，いわゆる談話認識（*speech recognition*）は，連続的な音声形式を分析したり，さらに，さまざまな方言，声の高さ，背景の騒音を克服しなければならない。

自動談話認識の質の尺度は，人間の談話認識である。こうして次のような要求が，自動談話認識のシステムに出される。

2.1 自動談話認識の要件

▽ 話者に依存しないこと

話者の発音が，声の高さ，方言，スピードなどの点で，異なっているとしても，システムは，さまざまな話者の自由闊達な談話を克服しなければならない。

▽ 分野に依存しないこと

システムは，内容の如何にかかわらず，話しことばを書きことばに転換することができなければならない。

▽ 現実的な語彙

認識可能な語形の数は，普通の話者のそれに相応しなければならない。

▽ 強 靭 さ

話しことばの中断，短縮，間延びがあったとしても，システムは，意図された語形を推論することができなければならない。

今日の談話認識システムは，その中ではただ限られた対話しか意味をなさない何らかの分野をあらかじめ設定することで（例えば列車の案内），ある程度話者に依存しないレベルに到達している。この利用分野の内容的制限についての知識は，── 文法的知識と組み合わせて ── もっとも可能性の高い語形を推論するのに用いられる。

しかしながら，こうした談話認識システムの語彙は，依然として 1,000 語形足らずの状態である。それに対し普通の話者はおよそ 10,000 単語を使用する。これはドイツ語では，およそ 100,000 語形に相当する。さらにまた平均的話者の受動的語彙

は，3ないし4倍多い。

この難点にもかかわらず，自動談話認識は，目下世界中で，集中的に，多額の費用をかけて研究されている。その理由は，口述筆記の方が，タイプ入力より，ずっと簡単（ユーザフレンドリ）だからである。この実用上の目標は，電子秘書から，電話による自動列車案内，さらには「移動通訳機」にまで及ぶ。これは，ドイツ語ないしは日本語を中に吹き込むと，（小さなスピーカーから）英語の[2)]翻訳が出てくるというポータブル・コンピュータである。

今日の音声談話認識システムが，音波の解釈の際に，文法の知識や専門分野に関する知識を，きわめて著しく中に含めていることは，それらの助けでともかく一つの成果に達しようというための非常処置とは決してみなされない。むしろこの戦略は，人間の状況に対応している。人間もまた，使える情報をすべて，話しことばの解釈の際に，投入しているのだから。

しかしこのことは，談話認識の課題が，話しことばの電子メディアへの転換以上でも以下でもないという事実をゆるがせにするものではない。電子メディアこそが，当然のことながら，辞書，形態論，統語論，意味論，語用論といったコンピュータ言語学的分析の本来のメディアである。

別のことばで言えば，電子的に記憶されたことばのコンピュータ言語学的分析は，他の言語メディアに依存せずに行われる。一方，こうしたコンピュータ上でのことばの一般的・抽象的分析の質や効力が高くなればなるほど，それは視覚的および聴覚的信号認識の基礎として，より有用なものとなる。

3. 電子メディアの技術的利点

言語学に基づく方法論の投入がなくても，電子メディアは，他のメディアに対して，きわめて本質的な利点を持っている。コンピュータ上での電子処理の可能性は，今まで印刷メディアのかたちでしか存在しなかったテキストが，なぜ電子メディアに転換され，今やCDで買うことができるのか，その理由が答えである。例えば，

▽ 古典ギリシア語の全テキスト
古典ラテン語の全テキスト

シェークスピア全集
ブリタニカ百科事典
ブロックハウス / ヴァーリッヒの辞典

例として10巻の事典の印刷版とCD-ROMの電子版の利用を比較してみよう。CD-ROM上で見たいテキスト箇所を探す際のスピードと快適さに有利性がある。書架から何巻かを取り出し正確なページを探す代わりに，CD-ROMの場合は，キーワードを入力するだけでよい。

適切なソフトウェアにより，主記載項目を探すだけでなく，記載項目中のキーワードの全出現箇所を見つけることができる。そして最後に，ワードの組み合わせを検索することができる。例えば，「画家」，「ベネチア」，「16世紀」というワードが40語の長さ以内に現れるすべての箇所を検索することができる。

この電子ベースの検索方法は，辞典の利用あるいはシェークスピアやギリシア・ローマの古典テキストに関する学術研究の際に，実際的な効用を持つばかりではない。法律データベースの助けで訴訟の準備をしたり，めったにない病気のコンピュータ支援診断あるいは特別な薬品の選択の際にも，電子データベースは，従来の著作物やカード・ボックスよりはるかに抜きんでている。

タイプライタに対して，コンピュータは，テキストを電子的に校正したり，別のファイルにコピーしたり，編集したり，サイズ変更したりする可能性を提供する。この理由から，今日出版されるたいていのテキストは，本来的には，電子の形式でつくられ，ようやく最後の段階になって，本や新聞印刷といった二次的メディアに転換される。

例えば日刊新聞を観察してみよう。以前は，個々の記事は，機械的な植字機で，一つ一つの文字からなる鉛の組み版の中で組み立てられた。日刊新聞の内容は，印刷プロセスの終了後再び壊されるか溶かされる印刷プレートのかたち，および紙の上に印刷された新聞冊子のかたちでのみ存在した。

編集部が報道部の原稿を引き受けようとする場合，原稿は一字一字オリジナルから写し取らなければならなかった。印刷の開始前に，ぜひとも採用したいトクダネが舞い込むと，新しい原稿のための余地を空けるために，印刷プレートを手で組み替えなければならなかった。

今日，新聞の内容は電子の形式で存在する。報道部の原稿は，もはや紙の上で供

給されるのではなく，電話を経由して，モデムの助けで電子メディアに再構成され得る形式でやって来る。電子の形式での新聞の出版は，任意に変形したり，コピーしたり，編集したりすることができる。その際，これらの版のどれでも，自動的に印刷することができる。

3.1に，ある日刊紙の短い記事を掲げる。これは，ある出版社の計算機に記憶されたかたちのものである。このテキストは，植字機のための特有の制御コードを含んでいる。

3.1　制御コード付きの新聞テキスト

```
0509636
     /  otagD22801P1008501271738otagotag
<01001> <SB15.HO80.HX2,3.DA2.SA51.SG8> <ef> politik
- panorama windelen - vd++ - otag<001,0003>
<01002> <002,0006> <01003> <sb14>  Heinrich
Windelen, ++ <mp> <SA50> <SG8> <SW> <SK> <DZ> <ef>
Bundesminister f}r <01004> Innerdeutsche++
Beziehungen, sieht Anzeichen <01005> f}r eine
Einigung zwischen Bonn und++ <01006> Ostberlin in
der umstrittenen Frage der <01007>
DDR-Staatsb}rgerschaft++ . Im Hessischen <01008>
Rundfunk meinte der CDU-Politiker, ohne <01009++>
auf Einzelheiten einzugehen, es verdichteten <01010>
sich die Indizien daf}r++ , da~ die SED-F}hrung
<01011> offenbar nicht mehr auf einer "vollen
Aner++<-> <01012> kennung" bestehe, sondern sich
mit einer <01013> "Respektierung++" durch Bonn
zufrieden geben <01014> k|nnte.<014,0042> <014,0042>
```

このテキストが，さてどのようにして例3.1の場所にやって来れたのだろうか。言語学者は日刊紙に，アクチュアルな情報のためではなく，それが言語の時代証言であるがゆえに，関心を寄せる。日刊紙は，今日，本来的に電子の形式で存在するので，言語学者にとって，電子的に記憶された新聞出版の中の任意の部分を入手することは，実際上単に法律の問題（版権）だけである。

許可がある限り，もはや必要なのは，印刷用テープのコピーを入手して，自分の

コンピュータに再生することだけである。その後は情報を任意に処理できる。例えば，一定のテキスト箇所を取り出してコピーし，3.1のように，他のテキストに，はめ込むことができる。さらに別の可能性は，制御コードを取り除いたり，解釈したりすることである。

3.2 「プレーンな」新聞テキスト

```
05.09.86
politik - panorama windelen
Heinrich Windelen, Bundesminister fuer
Innerdeutsche Beziehungen, sieht Anzeichen
fuer eine Einigung zwischen Bonn und
Ostberlin in der umstrittenen Frage der DDR-
Staatsbuergerschaft. Im Hessischen Rundfunk
meinte der CDU-Politiker, ohne auf
Einzelheiten einzugehen, es verdichteten
sich die Indizien dafuer, dass die SED-
Fuehrung offenbar nicht mehr auf einer "vollen
Anerkennung" bestehe, sondern sich mit einer
"Respektierung" durch Bonn zufrieden geben
koennte.
```

テキストが，印刷された新聞として別に存在しない場合は，テキストの文脈を通して，制御コードの解釈を行うこともある。例えば，3.1の“Staatsb}rger-schaft”は，明らかに *Staatsbürgerschaft* のことであり，“k|nnte”は，明らかに *könnte* のことである。

今日なお世界的に，各国のみならず，実際上，各印刷業者が，制御コードのために独自の規約を持っているという問題が存在する。絶えず変わる制御の規約を解釈することは面倒で，時間の無駄であるので，**国際標準化機構**（ISO）によって，いわゆるSGML標準規格が開発された[3)]。

3.3 SGML: standard generalized markup language.

電子版テキストに標識をつけるためのISO標準規格の一つで，テキストの送り手および受け手双方に，テキストの構造を識別させ

ることができる。(すなわち標題，著者，ヘッダ，パラグラフなど)

Dictionary of Computing, S. 416 (Illingworth et al. 1990)

SGML 標準規格は，ヨーロッパにおいても，それゆえドイツにおいても評価されており，時とともにますます広く普及しつつある。というのも，この標準規格の規約に沿った電子テキストは，その制御コードを，他のすべての SGML ユーザが，自動的に解釈できるからである[4)]。

コンピュータに記憶されたテキストは，ユーザの個人的意図と要求に応じて電子的に変更することができる。例えば，3.2 の「プレーンな」新聞テキストは，エディタによって次のような LaTeX での処理用に加工することができる。

3.4 テキストの LaTeX-整形

```
\documentstyle{artikel}
\begin{document}

\noindent
05.09.86\\
{\bf Politik:} - {\it panorama windelen}\\
Heinrich Windelen, Bundesminister f\"{u}r
Innerdeutsche Beziehungen, sieht Anzeichen
f\"{u}r eine Einigung zwischen Bonn und
Ostberlin in der umstrittenen Frage der
DDR-Staatsb\"{u}rgerschaft. Im Hessischen
Rundfunk meinte der CDU-Politiker, ohne auf
Einzelheiten einzugehen, es verdichteten
sich die Indizien daf\"{u}r, da{\ss} die
SED-F\"{u}hrung offenbar nicht mehr auf
einer "vollen Anerkennung" bestehe, sondern
sich mit einer "Respektierung" durch Bonn
zufrieden geben k\"{o}nnte.
\end{document}
```

LaTeX は，コンピュータ上での組み版用のプログラミング言語として，D. Knuth

によって開発された TEX の簡易ヴァージョンである。3.4 が LATEX プログラムによって送られると，コンピュータは次のような文字イメージを出力する。

3.5 テキストの TEX で加工された版

05.09.86
Politik: - *panorama windelen*
Heinrich Windelen, Bundesminister für Innerdeutsche Beziehungen, sieht Anzeichen für eine Einigung zwischen Bonn und Ostberlin in der umstrittenen Frage der DDR- Staatsbürgerschaft. Im Hessischen Rundfunk meinte der CDU-Politiker, ohne auf Einzelheiten einzugehen, es verdichteten sich die Indizien dafür, daß die SED- Führung offenbar nicht mehr auf einer "vollen Anerkennung" bestehe, sondern sich mit einer "Respektierung" durch Bonn zufrieden geben könnte.

3.4 と 3.5 では，ただごく簡単な LATEX のコマンドが表示されている。例えば，強調（`{\bf }`, *bold face*）と斜体（`{\it }`, *italic*）である。さらにウムラウトやエスツェットおよび右方禁則の印刷イメージを取り扱うことができる。この場合プログラムは，行末での語分割を自動的に実行する。

さらに，章やセクションの標題の自動処理，内容一覧やインデックスの自動作成およびその他もろもろがこれに加わる。とくに数式の表示の場合に，TEX および LATEX は，きわめてパフォーマンスが高い。

1984 年のその導入以来，TEX および LATEX は，学術文献や雑誌の出版に，ますます多く使用されている。その際，研究者は，自分の論文や著書をコンピュータ上で書くばかりでなく，自分で形式を整え，印刷スタンバイの形で，出版社に引き渡すのである。電子メディアを経由した出版は，伝統的な活版の本より，コストがずっと有利であるばかりでなく，多くの実際的な利点を持っている。とくに著者が

製作に直接タッチできること，および植字工の仕事から校正読みがなくなることなど。

テキスト箇所の高速なコンピュータ支援による検索やデスクトップ・パブリシング（DTP）と並んで，電子的に記憶されたテキストは，有用な言語学的分析の可能性をも提供する。例えば3.2のテキストは，わずかのステップでアルファベット順の単語リストに変形することができる。

3.6　テキストのアルファベット順語形リスト

05.09.86	Windelen	koennte
Anerkennung	auf	mehr
Anzeichen	auf	meinte
Beziehungen	bestehe	mit
Bonn	dafuer	nicht
Bonn	dass	offenbar
Bundesminister	der	ohne
CDU-Politiker	der	panorama
DDR-Staats-		
buergerschart	der	politik
Einigung	die	sich
Einzelheiten	die	ohne
Frage	durch	sieht
Heinrich	eine	sondern
Hessischen	einer	umstrittenen
Im	einer	und
Indizien	einzugehen	verdichteten
Innerdeutsche	es	vollen
Ostberlin	fuer	windelen
Respektierung	fuer	zufrieden
Rundfunk	geben	zwischen
SED-Fuehrung	in	

3.6のような単語リストは，すべての個々の語形の出現を数え，こうしてテキスト中の語頻度についての統計的研究の基礎を与える。一方で，各語形が一度しか現れない（そして大文字と小文字の区別がなされない）単語リストも同様に容易に作成することができる。この第二のタイプは，例えば，語彙の分類に適切なものである。

これまで述べてきた，大きなテキスト・データベース中の単語や語句の自動検索，

自動誤り分析（単語リストを参照しての「スペル・チェッカー」）といった処理，制御コードを使用した整形，テキストのアルファベット順語形リストへの変形などは，電子メディアにおける記号処理の純粋に技術的な操作である。それらは，決して言語学的なコンセプト，理論あるいは方法論に基づくものではない[5]。

非電子的な方法（活版，カードケース，ページ繰りや通読のかたちでの大きな文書の検索）と比較して，こうした電子的な方法は，きわめてスピーディに，正確に，快適に処理することが可能である。これらはテキストを扱う実際の仕事を軽減するだけでなく，言語学的分析のための価値あるデータを提供する。（語形のアルファベット・リスト，語形の頻度に関する統計的記述，大きなテキスト中の語形の二つ，および三つの組み合わせ――いわゆるトリグラムなど）

しかし同時にまた，はっきりとした限界もある。それは，技術に基づく方法が，純粋に文字を拠り所にしているということである。語形，文構造，およびそれに立脚した内容の文法的分析は，こうした技術の範囲を超えるもので，言語学の分野に位置づけられる。

4. 文法の構成部門

言語学の方法によって，どの分野で，電子的テキスト処理の本質的な改善が達成されるのであろうか。この問いに答えるための最初の基礎として，以下に文法の構成部門とその働きを述べることにする。

その際，言語学の内部で，三つの異なる文法分析のアプローチがあることを考慮しなければならない。つまり (a) 伝統文法，(b) 理論言語学，(c) コンピュータ言語学である。これら三つのアプローチは，その

1. **方　法**
2. **問題設定**
   （つまり記述的ないしは説明的目標）および
3. **応　用**

に関して区別される。

文法の構成部門を述べる前に，言語学内部での，この三つのアプローチの図式的

な比較を始めよう。

4.1 言語分析の三つの異なるアプローチ

▽ 伝統文法

伝統文法は，方法論から見て，タクソノミック（記述的—分類的）な指向を持っている。

その目標は，個別的言語現象のできるだけ完全な収集と分類である。とくに言語の規則性と例外を記述する。

応用面から見ると，それは語学の授業から由来するものである。

コンピュータ言語学にとって，伝統文法は，その豊富な経験的資料のために，大いに興味あるものである。

▽ 理論言語学

理論言語学の方法論は，論理—数学的である。つまり，すべての適格な言語構造のみを派生することができる形式規則体系を定式化する。これは，伝統文法に対して，明示的な仮説形成という方法論的利点を持っている。ただしこれは，単なる理論的なものである。というのも，現実の多量の資料で形式規則体系を検証することは，紙と鉛筆では実際には不可能だからである。

理論言語学は，相変わらず多くのさまざまな学派に分裂しているけれども，共通の説明的目標は，人間の言語能力の形式的記述ということである。ちなみに，この場合，言語運用（*Performance*）の側面は排除される。

応用の試みは，心理学における説明的モデルから，学校の語学授業にまで及ぶ。

コンピュータ言語学にとっては，とくに形式的言語階層と複合性についての研究が重要である。

▽ コンピュータ言語学

方法論的には，コンピュータ言語学は，自然言語現象のできるだけ完全な分類という伝統文法の目標と，理論言語学の論理—数学的アプローチを結び付ける。ただしこれに重要な革新が加わる。つまり，明示的な仮説は，パーザーとしてインプリメントされた文法によって表示され，自動的に大量のデータによって検証され得る。

コンピュータ言語学の記述的および説明的な**究極目標**は，自然語を用いた情報伝達のモデル化である。この目標に至る途上で，自然言語の完全な形態論的，語彙論的，統語論的，意味論的，語用論的把握が，実用という機能的枠内で，なされなければならない。

この目標に到達することにより，自動的言語処理の適用に，広大な可能性が開かれる[6)]。

すでに述べた言語学の諸学派は，その方法，目標，応用がさまざまであるにもかかわらず，文法を，音韻論 (*Phonologie*)，形態論 (*Morphologie*)，辞書 (*Lexikon*)，統語論 (*Syntax*)，意味論 (*Semantik*) および補足的分野である語用論 (*Pragmatik*) 各部門に分割するという共通の基礎を持っている。ただし，これらの部門の価値の置き方や学問的取り扱いは，言語学の諸アプローチにおいてさまざまである。

4.2 文法の構成部門

・音 韻 論

言語音声の科学

理論言語学において，音韻論は，一種の基礎学問として中心的な役割を演じている。そこでは，言語分析の普遍原理（弁別的素性，形式的規則装置）が経験的に提示される。その目標は，(a) 歴史的変化（音声変化）あるいは (b) 発音の共時的交替（例えばドイツ語のいわゆる「語末音硬化」[訳注2)]）を，規則体系の形で，できるだけ一般的かつエレガントに記述することである。

それに対しコンピュータ言語学においては，音韻論は，たいていの場合，従属的な役割を演じる。場合によって導入できそうな唯一の分野は，自動談話認識である。ただしこの分野は，今日，(音韻論ではなく) 音声学 (*Phonetik*) を援用して行われている。音声学は，(a) 発声，(b) 音響，(c) 聴音の各プロセスの構造を研究する。音韻論に対し，音声学は，文法の部門に数えられない。

・形 態 論

言語の語形に関する学

形態論は，例えば（一例としてラテン語の）学校文法に見られるように，伝統文法の主要分野である。それは，ある言語の単語を，品詞に基づいて分類し，活用，派生，合成に関して各語形を記述する。

コンピュータ言語学においては，いわゆるコンピュータ形態論が，**自動語形認識**の課題を持つ中心的な領域である。これはオンライン辞書と形態素解析プログラムに基づいて行われる。自動語形認識は，言語学に基づく他のすべての自動テキスト分析操作の実際的な前提である。

・辞　書

言語の単語のリスト化

ある言語の単語のできるだけ完全な収集と整理は，辞書学（Lexikographie）と語彙論（Lexikologie）の分野に入る。辞書学は，語彙項目の辞書記載法と構成の原理を扱い，実用に向けられた言語学の周辺領域である。語彙論は，ある言語の語彙を，その内部的な意味構造を考慮して研究するもので，伝統的言語学にその故郷を持っている。

理論言語学に関しては，60 年代半ば以来，辞書に関する関心が絶えず高まってきていることが特筆される。こうした研究の特徴的な傾向は，複合的語句のますます多くの統語的・意味的特性を，その中に含まれる単語の辞書的特性から派生するというものである。その成果は，個々の単語の広範な形式的記述であり，これを統語現象の解明に役立てようとしている。

コンピュータ言語学においては，自動語形認識の際に，オンライン辞書が，形態素プログラムと組み合わせて使われる。その目標は，最小限のメモリ容量と速いアクセスで，最大限可能な完全性を手に入れることである。自動語形認識の枠内で新しい辞書を作成することと並んで，オックスフォード英語辞典（現在，電子の形式で存在する）のような伝統的な辞書の知識を，自動テキスト分析に役立てようとする関心も高まっている。（「辞書採掘」）

・統 語 論

語形の文法的に適正な組み立ての記述

理論言語学（生成文法）においては，統語分析の目標は，ある言語における文法的適格性の記述である。つまり形式的規則を用い，これは，ある言語のすべての，かつ適格な表現のみを生成する（産み出す）ないしは認識する。豊富な形式的可能性の中から，持続性のある正確な記述を見つけるために，この場合第一に，いわゆる普遍性に基礎を置いた人間の言語能力の記述が追求される。

コンピュータ言語学は，一方で，自然言語のための形式文法の生成能力を，実際に多量のデータで有効に検証するという技術的前提を提供する。他方，有用な自動構文解析に対する大きな実用的需要に直面して，過去30年間に，独自の，コンピュータ指向のシステムが開発され，これらは，理論言語学の統語論体系に，多かれ少なかれ直接に影響を与えてきた。

・意 味 論

言語表現の逐語的意味の分析

理論言語学においては，意味論の課題は，(異なるないしは同一の)「深層構造」を用いて統語的曖昧性やパラフレーズを記述することから，論理式を用いて真理条件を論理—意味論的に表示すること（例えばモンタギュー文法）に及ぶ。その場合，語意味論という部分領域が，単語ないしは語形の意味分析を扱う。一方，文意味論は，複合的語句の意味が，その部分の意味およびその組み立ての様態から，いかにして派生するかを記述する。(フレーゲの原理)

言語表現の意味分析は，中でも，単数および複数（量化詞），連接（*und*），選言（*oder*）といった論理学的記述，主語と目的語による動詞の補足（格と結合価），形容詞と副詞による名詞句および動詞句の修飾，節の従属関係，その他もろもろを含む。コンピュータ言語学においては，意味分析の問題点は，プログラミング言語の解釈から，データベースの整合性，さらには概念ベースによるインデックス化の手順や機械翻訳における曖昧性除去の手順にまで及ぶ。

・語 用 論

言語表現の使用の理論

これまでの部門（音韻論，形態論，辞書，統語論および意味論）が言語表現（語形や文）の構造的特性を扱うのに対し，語用論は，言語表現の使用の際に，これらの構造的特性が，ある発話コンテキストにおいて，どのように作用するかを研究する。したがって厳密には，語用論は，文法の部門に属するのではなく，(i) 言語表現の構造分析（文法），(ii) (発話および解釈の) 文脈の記述，(iii) 言語とコンテキストの相互作用の分析を包括する。

人間と人間の，あるいは人間と機械の間の自然言語による意味伝達を，理論的かつ実際的にモデル化しようとする場合，この三つの下位部門を持つ語用論は，決し

て欠かしてはならないものである。言語表現の使用には，指示（つまり言語表現を話者によって意図された対象に関係づけること），指標的表現（代名詞，時間および場所の副詞，動詞活用）の解釈，表現を作り出す際の代名詞や語順の修辞的に正しい使用，そして言葉どおりではない用法の解釈（例えばメタファー）が含まれる。

理論言語学において，語用論は，たいてい，モデル理論的（論理学的）意味論あるいはいわゆる発話行為理論（Sprechakttheorie）の枠内で取り扱われる。コンピュータ言語学においては，語用論は，（例えば対話システムや機械翻訳システムでの）生成部分を修辞的に正しくインプリメントする際の実際的な諸問題をきっかけに関心を見出している。

これまで述べてきた文法の各部門の分割は，自然言語の異なった構造的側面，つまり音声（音韻論），語形（形態論），単語（辞書），文（統語論），意味（意味論），使用（語用論）に合わせたものなので，これは，伝統的言語学，理論言語学，コンピュータ言語学にとって同様に通用する。

5. コンピュータ言語学の究極目標

今日まで，理論言語学の諸文法は，ただ，そのエレガントさ，数学的集合の大きさ，扱える資料の量，あるいは心理学のテストとの互換性の面から評価されてきた(そしてされている)。これに対しコンピュータ言語学の方法は，さらに，形態，統語，意味派生の形式的構造が，簡単で速い誤りの検出や拡張を許す見通しのよいプログラムの流れのためのよい基盤を提供する，ということを必要とする。その他，人間のことばによる意志疎通をコンピュータ上にモデル化するためには，有用な語用論的解釈のための文法分析が，分析の際にも，生成の際にも，もっとも効果的なのだという点が欠かせない。

それゆえ学問的見地からは，コンピュータ言語学的アプローチにおいて本質的なことは，可能な応用（確かに重要でおもしろいが）ということより，むしろその新しい方法論にある。コンピュータ言語学の究極目標，つまり人間の言語使用をコンピュータ上にモデル化すること（4.1参照）は，それが機能的・発見的見方を強制するため，方法論的にきわめて大きな意味を持つ。

理論言語学の最近の歴史は，その提唱者が，現在ある部門と形式装置を，そのためにはそれらがまったく開発されていない諸現象に適用するという誤りに，よく陥ることを示している。そしてこれは，記述装置がこうした拡張を表面上許容するように見えるという理由からに過ぎない。こうした例は，文献に数多く見つかる。例えば，意味論的現象を統語論で扱ったり，語用論的現象を意味論で扱ったり，形態論的現象を統語論で扱ったりなど。これらはくりかえし袋小路に陥り，再び脱出するのに数十年かかった（またかかっている)。こうした危険は，一貫性のある作用能力を持った総体的コンセプトによってのみ，実際に追放することができる。

コンピュータ言語学の究極目標は，現実的な自然言語コミュニケーションのモデルを開発しインプリメントすることが，原理的にそもそもどの程度まで可能なのか，という問題に至らせる。一つのアナロジーによってこの問題に答えたい。

コンピュータ言語学の今日の状況は，多くの点で，機械による飛行の発展に対応している。何百年も人間は，どのように飛ぶのかを理解するために，スズメや他の鳥たちを観察してきた。そして，できるだけ似た方法で，空中に舞い上がることを試みた。

そして翼をばたつかせるのではうまくいかないことが明らかになった。このことは，人間が飛行することは原理的に不可能である，と断言する口実に好んでされた。しばしば「人間が空を飛ぶことを神が望んでいたとしたら，神は人間に翼を授けたであろう」という敬虔なことばとともに。

今日，飛行は人間にとって当たり前のことになった。さらにその間に，スズメはジャンボジェットと同じ理論的原理，つまり「エア・フォイル」(揚力翼面）の原理で空中にとどまる，ということがわかった。つまり，スズメの飛行とジャンボジェットの飛行は同一の原理で行われている，とまとめられる抽象のレベルが存在する。

自然言語のコミュニケーションをコンピュータ言語学においてモデル化する際にも同様に，適正な抽象のレベルに立脚した適正な原理が問題となる。その場合当然のことながら，レベルを低く設定し過ぎたり，高く設定し過ぎたりする危険が存在する。例えば自動券売機に見られるような閉じた信号体系は，確かにモデルとしては不適切であろう[7]。

しかしまた，自然言語のコミュニケーションのモデル化を，はじめから，素朴な知った顔のイメージで，不合理であるとすることも，同様にナンセンスであろう。

例えば，「理解」という概念，これを追究すると，システムは『フィネガンズ・ウェイク』[訳注3)]を分析する際に，こと細かに分析に打ち興じなければならないが，この概念で取りかかる者は，ジャンボジェットから交尾行動や卵の世話を期待する人と同じような地点にいるのである。

最後に飛行機製作の歴史からの第二のアナロジー。人間が，複葉機，プロペラ機，ジェット機で，飛行原理の理解を次第に改善してきた後，今日では再び，自然の飛行プロセスを重点的に分析し，その素晴らしい能力を把握して，より静かでより効率的な飛行機の製造にうまく取り入れようとしている。

この例が示しているのは，コンピュータ言語学におけるさまざまな理論的・技術的解決の試みは，人間の言語能力の分析についての関心の欠如を含意するものでは決してないということである。むしろ，自然言語の原理的なモデル化がなされ，さまざまな応用の中でその真価が実際に示された後になってようやく，言語コミュニケーション過程におけるすぐれて人間的なものの研究が，実際に意味のあるものになるということなのである。

原注

1) これについては D. McClelland 1991 参照。

2) 利用の前提として想定されるのは，会話のパートナーであるドイツ人と日本人が，英語の受動的な知識を持っていることである。こうして，聞き手が話し手の言うことを理解できるだけでなく，話し手も機械の自動翻訳をチェックすることができる仕組みになっている。

3) SGML のテーマについての詳しい説明は，Herwijnen 1990 にある。

4) この間に，非常に強力な SGML の簡易版として，TEI 標準が提案された。TEI は，SGML を部分特化したもの（サブセット）で，*text encoding initiative* の略である。

5) しかし，たとえばデスクトップ・パブリシングの枠内での，行末での単語の自動分割のような単純な問題においてすでに，技術と言語学が，自動テキスト処理の際に密接に結びついていることがわかる。

6) 当初から実効ある処理を目標にして開発された文法形式論の一つは，左方連想文法 (Linksassoziative Grammatik = LAG) である。LAG の複合的特性は Hausser 1992 に述べられている。

7) ここで問題になっているモデル化の際の決定的なポイントは，自然言語コミュニケーションに特徴的な多面性が保たれているということである。── つまり，

同一の表現が，きわめてさまざまな発話コンテキストにおいて，有意義に投入されうるという事実である。

参考文献

Illingworth et al. (Hrsg.) (1990) Dictionary of Computing, Oxford University Press, Oxford.

Hausser, R. (1989) Computation of Language, Springer-Verlag, Symbolic Computation: Artificial Intelligence, Berlin–New York.

Hausser, R. (1992) "Complexity in Left-Associative Grammar," Theoretical Computer Science, Vol. 103: 283–308, Elsevier.

Herwijnen, E. van (1990) Practical SGML, Kluwer Academic Publishers.

Hutchins, W. J. (1986) Machine Translation: Past, Present, Future, Ellis Horwood Lmt., Chichester.

McClelland, D. (1991) "OCR: Teaching Your Mac to Read," MACWORLD, November 1991: 169–175.

〔訳注〕

1) 本稿は Hausser, Roland: Aufgaben der Computerlinguistik. In: LDV-Forum Bd. 10, Nr. 2, Jg. 1993. S. 63–77 の全訳である。LDV-Forum はドイツ言語情報処理学会（Gesellschaft für Linguistische Datenverarbeitung = GLDV）の機関誌。著者の Hausser 氏は，現在エアランゲン・ニュルンベルク大学の教授（言語情報論 / 独文学科）。この論文では，コンピュータ言語学の現時点での到達点が明解かつ簡明に述べられており，広く専門外の人にも関心を惹く内容になっていると思う。なお，この論文の発展ともいえる Grundlagen der Computerlinguistik という書物が近刊予定である。

2) Auslautverhärtung. 語末またはつづり末での子音の無声音化（例：Geld, Lob, Tag）。

3) "Finnegans Wake" (1939). アイルランドの作家 James Joyce (1882–1941) の小説タイトル。完結していない小説の最後の部分が，冒頭にあらわれたり，さまざまな言語をマージするなど斬新な手法が使われており，難解で知られる。