OCR（光学文字読み取り装置）によるコーパス作成

舟阪　晃

1　まえがき

言語の研究や、言語と関わりのある仕事に携わる人間にとって、コーパスの存在はこのうえなく重要であるといえる。とくに、対象言語が母国語でないときは、質の良いコーパスを手近にもつか、もたないかは、その成果に致命的な影響をあたえる。

コーパスといえば、アメリカ英語についてはBrownコーパス、イギリス英語ではLOBコーパスが、思い出されるが、かならずしも簡単には入手できなかったり、収録されているデータの配列が納得できなかったり、年代的にすこし古くなっている、など問題もある。

わが国において、私の知るかぎり、最初にBrownコーパスを有効に、組織的に使用した研究書は、梶田優（68）：*A Generative-Transformational Study of Semi-Auxiliaries in Present-Day American English*（三省堂）であると思うが、それ以後、研究論文などにも、おりにふれて、このコーパスについての言及を眼にするようになった。

いわゆるコーパスという意図で編集されたものではない文法書を、しばしば、コーパスとして援用していることもある。たとえば、Jespersenの*Modern English Grammar*や、Randolf Quirk他の*A Comprehensive Grammar of the English Language*などは、その好例であろう。また、勝俣銓吉郎著『英和大活用辞典』を参考にしていることもある。

このように考えてくると、われわれがいかにコーパスを必要とし、にもかかわらず、いかにコーパスにめぐまれていないかがわかる。ひとつには、自

ら手を汚すことなく、外国の学者の成果にいつまでも依存しているのは、工業技術の面とおなじく、日本人のお家芸とはいえ、情けないといわざるをえない。また、他方、本の形になったものは、索引などが不完全なことが多く、検索に時間がかかりすぎて、よほどの問題でないと、検索に時間と労力をさく気持ちになれない。ハイテク時代に、いつまでも、紙製の辞書や参考書と悪戦苦闘するのは時代遅れというものであろう。地球上の植物資源の枯渇が心配されるおり、紙を使わないコーパスの、しかも日本人独自の作成が期待される。

前置きが長くなったが、コンピュータを活用した、われわれの手作りのコーパスを作る必要があり、その点から、問題点のひとつを考察しようというのが本稿の目的である。Merja Kytö 他編の *Corpus Linguistics, Hard and Soft* をみてみると、現代のハイテクを踏まえたいろいろの企画が出始めていることが感じられる。将来、Brown や LOB とならんで、「外大コーパス」とよべるような、英語に限らず、世界の諸語のコーパスができれば、社会に対する大きい貢献となろう。

2　コンピュータによるコーパス作成

コンピュータを活用してコーパスを作成するとき、技術的につぎの三点が問題とされてきた。

（1）入力の手間

（2）記憶容量

（3）検索速度

まず、（2）の記憶容量の問題は、ハードディスクや光ディスクの開発により、今や、問題ではない。つぎに、（3）の検索速度は、コーパスの有効な利用のために不可欠のものであるが、コンピュータの演算速度がどんどん早くなり、徐々に、優秀なソフトが開発されていく可能性が高いので、将来は明るいといえよう。最後に（1）の入力にかかる時間と労力は、いまだ解決されない大きな問題として残っている。

OCR（光学文字読み取り装置）によるコーパス作成

本稿では、入力のひとつの方式としてのOCR（Optical Character Reader）の実験的な試用を報告し、先行されている方々からのご教示をいただきたいと思う。

2．1　ハードとソフト

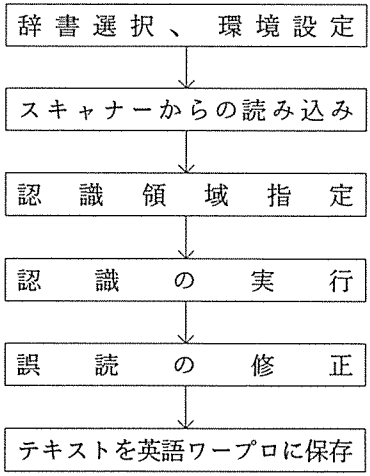
ここ一年間ほど、実験的に試用してきたOCRのハードとソフトの構成はつぎのようである。

（1）本体：　NEC 9801RA21
　　EMS:　MELCO EMJmkIII 2M
（2）HDD: ICM MC40S 40M
（3）スキャナー：　EPSON　GT6000
（4）ソフト：PCR-English

（バーズ情報科学研究所）

2．2　OCRの読み取り手順

文字認識の処理の流れは、概略、つぎのようになっている。



PCR-Englishのメニュ画面には、①イメージ、②認識、③テキスト編集、④辞書、⑤環境設定、があらわれる。

まず、④辞書を選択し、使用する辞書を指定する。たとえば、雑誌*Time*を読み取らせるときは、*Time*用の辞書を選択する。読み取る対象により、専用の辞書を使用したほうが読み取りの正確さが高くなる。なお、それぞれの辞書は、使用者が作ることになる。

つぎに、⑤環境設定では、原稿の文字間隔や文字の種類を指定し、スキャナーの機種決定、スキャナーの読み取り範囲の指定を行う。ちなみに、上記スキャナーの場合は、A4版が読み取り最大範囲となる。同時に、スキャナーの輝度や線密度も指定する。

ここまでは、準備の段階で、つぎに、読み取りの手順に入る。

①イメージの画面で、スキャナーの上においた原稿を読み取らせる。スキャナーは、文字を、文字としてではなく、イメージとして読み取る。A4で、約1分10秒ほどかかる。つぎに、罫線があるとそれを文字とまちがって認識することがあるので、罫線の消去を指示する。これが約30秒かかる。

つぎに、②認識を選択し、スキャナーが読み取ったイメージを文字として解釈させる。写真や図などがあるときは、それを避けて、文字の部分だけを領域指定して、読み取らせていく。辞書に依存する度合いを大きくすると読取りの精度は上がるが、読み取り速度は遅くなり、一方、辞書の利用を控えめにすると、速度は早くなるが、精度は落ちることになる。読取りの効率は、いかにして精度を落とさずに読み取り速度を早くするかにかかっている。

最後に、③テキスト編集の画面で、読み取られたディスプレイ上の文字と、原稿を照合し、訂正をし、最終的な仕上げをする。これは、われわれが肉眼で行わなければならない。原稿の紙質や活字の状態により、読取りの精度は大きく影響をうける。たとえば、雑誌*Time*のように良質の紙に印刷されている場合は、成績は比較的良好であるが、新聞の文字は、活字が紙ににじんで、間違いが多く、ここで扱っているソフトに関していえば、ほとんど実用にならないといってよいほどである。参考までに、実例をあとであげている。

2．3　読取りの精度

つぎに、手近かな原稿の読み取り精度を、1,000文字に対して、何文字の読み間違いがあったかを調べることによって、示してみる（下表）。
取り上げたのは、科学雑誌 *Discover*、雑誌 *Time*、*Newsweek*、それに、新聞 *The New York Times* の社説である。

Discover	28／1000
Time	30／1000
Newsweek	52／1000
New York Times	97／1000

科学雑誌 *Discover* は、今回の調査では、*Time* とあまり違いがないが、これまでの入力の経験からいうと、現在使用中のソフトとかなり相性がいいという感触をえている。誤読の28には、数字のゼロを文字の o に読み間違っているところが7ヵ所あり、また、イタリックの読み間違いが5カ所ある。一方、*Time* の集計した領域には、このような項目が含まれていなかったので、*Discover* の誤読の28はもう少し甘くみてもよいように思われる。

つぎに、*Discover* の読み取り例をあげておこう。下線の部分が間違いで、正解は、誤認文字のすぐあとの（　）のなかに、参考までに示しておく。

Discover（91年2月号）

We might, as the Department of Energy suggests, sn(a)ve oil by innA(fla)ting our automobile tires more fully, but then we risk a serious air shortage. And such a measure would just postpone th9(e) inevitable : the only workable long-term solution is the development of alternative fuels.

上の例の2行目の *inflating* の読み間違いは、一般的に、*fl*は、その2文字をかなり接近させて印刷するという習慣によるもので、このソフトは、すくなくともいまの時点では、そこまでは「賢く」ないといえる。しかし、徐々

にソフトを「教育」していくことは可能である。

雑誌*Time*は、今回調査した範囲では、かなり良い結果がでているが、単語間が、行によりかなりばらつきがあり、空間が狭い場合は、空間をあけないで読み取ってしまうという間違いが生じやすい。

Time（91年7月8日号）

Some u(o)f the fundamental images of the jhin(Am)erican gallery of national icons have receivc(e)d a dramatic reworking.

雑誌*Newsweek*は、文字の読み間違いも多いが、とくにめだつのは、単語間の空間を見落としてしまう場合が多いことである。下の例で、単語間のあるべき空間がないところには、参考までに、caretを入れておく。

Newsweek（91年7月8日号）

This was an S( )explosive allegation. P(F)inance Minister B(R)yutaro Hashimoto, an ambitious 7(r)ising powe 7(r) in Japan's rulin(g) Liberal Democratic Party, had explicitly denied/\( )/\such ( ) suggestions g(a)ll/\( ) week.

最後に、新聞*New York Times*の社説をとりあげるが、誤読率は、上にあげたごとく、97／1000となり、数字をみたかぎりでは、それほど悪くはないが、場所により、ため息が出てしまうような出来事栄えがみられる。その例を、下にあげておく。

New York Times（91年6月30日）

Justice Powell, no radical'(,) had reasoned that to permit eo(n)tf(r)eaties for S(s)uro(v)iving relatives at the penalty'(,) phase'( ) of capital trials/\( ) would be highly pref(j)udicial and '( )lead to arbitrary sentences. The o(0) hio caa(s)e tTh(u)rned oU

(u)t to be inapp?(r)opriate,( ).

以上読取りの実例をあげたが、つぎに、すこし視点をかえて、それぞれの雑誌の欄内の一行の文字数を調べ、誤読率との関係を考えてみる。

Discover、*Time*、*Newsweek* の三誌は、A4の紙面を3欄にわけているが、それぞれの欄ごとの文字数は、下のようになっている。

	平均	最大	最小
Discover	29.6	33	25
Time	34.2	40	30
Newsweek	35.6	40	32

誤読率は、平均文字数と関係があることがわかる。平均文字数が多くなると、単語間の空間が正しく認識されなくなり、誤読が増えることになる。ただ、この種の問題は、資料ごとの専用の辞書の精度を上げ、環境設定を適切にすることにより、かなり改善していくことが可能であると思われる。

2．4　PCR-English の評価

今回使用したソフトの仕事ぶりをどのように評価するかは、使用者の要求水準に左右されることになろう。文字読み取りのためのソフトと銘打って、かなり高い代価を要求する以上、100%に近い成績でなければならないとする向きには、かならずしも満足できる成果ではないかもしれない。一方、本来、人間の頭脳がやっている複雑な文字認識の仕事が、コンピュータにもできたということで感激する人にとっては、刮目すべき成果と思われるであろう。私の個人的な評価は、どちらかといえば、上記の後者の立場に近いが、さらに、私にとって幸いなのは、このソフトを使うことにより、マニュアル入力で要求される目の疲労を減少させることが可能であるという点である。

マニュアル入力では、通常、下の手順が必要となる。

（ｉ）原稿を見ながら入力

（ii）必要ならスペルチエックの使用

（iii）原稿とディスプレイ上の文字の照合

これにたいして、ソフトによる読取りの場合は、（i）の段階から解放されることになる。しかも、この段階が、一番目に対する負担が大きいことは周知の事実であるといえよう。もちろん、ソフト使用の場合、（iii）の段階では、マニュアル入力のときよりは、多くの負担を要求されることになるが、（i）の負担よりは少ないのがふつうである。ただし、誤読率が、*New York Times* のそれぐらいになると、むしろ、マニュアル入力を選択したほうが得策であるといえる。

結論として、誤読率の低かった原稿に対しては、この種のソフトはそれなりの働きをするといえよう。辞書を精密にしたり、環境設定を適切にすることにより、ソフトを「教育」すれば、効率を高めることも可能である。

2．5　コーパスの保存

コーパスの保存をどのようにするかは、検索の効率にも影響するので、重要である。Brown コーパスや LOB コーパスともに、内部的な分類が、階層をなしていなかったり、分類されたテキストどうしが相互排除的でなかったり、不満が多いが、いざ代案をだそうとすると、容易な問題ではないことに気が付く。私の場合、今のところ、読み取りソフトの性能を調べている段階で、保存の方式については、まだ考えていないといってよいほどであるが、一般的にいって、次の点は、考慮されなければならないであろう。

（i）時代的分類

（ii）地域的分類

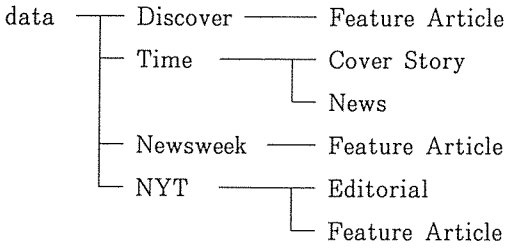
（iii）ジャンルによる分類

（iv）文体による分類

（v）使用域（register）による分類

ちなみに、私は、現在のところ、雑誌や新聞に限っているが、資料をハードディスクに入力するときは、下のような要領で行っている。これは、確定

的なものではなく、将来、より好ましいかたちにしやすいように考慮しながら、実験的に入力している。



同時に考慮されねばならないのは、いわば生の資料を収録するだけではなく、文法的なタグ（tag）をどのように付与するか、という問題である。たとえば、生の資料と文法的なタグをどのように共存させるか、さらに、文法的な記述法が複数ある場合そのどれを採用するか、など問題は多いが、これらは、本稿の目的から逸脱するので、稿を改めねばならない。

3　あとがき

言語学の観点から、コーパスの機械的処理のごく限られた一面について論じてきたのであるが、もっと一般的な観点からも、年々増えることはあっても減ることはない情報やデータをどのように保存、活用していくかは、大きな問題であるといえる。いわば、質の良いソフトをいかに身近に準備できるかということは、われわれの文化の成熟度の指標であるともいえよう。日本のハイテクは、主に、ハード面において大きい成果を上げてきているが、今や、ソフト面に注意を向ける時期がきているといえる。（91年10月31日）