るのは, この構想中のネットワーク版の SMART-GS が使われるようになったときであろう.

3.5 Reading unreadables：読めないものを読む

人文学における文書研究の最大の目標の一つは,「Reading unreadables：読めないものを読む」をいうことであろう. 通常, 文書とは苦もなく読めるものだと前提されている. 一つの文書が何通りにも読めるというのは, その意味だけであって, それがどういう文字の並びであるかには, 人によって意見の相違などない, というのが, 多くの人々の常識であろう.

しかし,「テキスト分析」「テキスト理解」についての, この前提は, 人文学の研究においては必ずしも成り立たない. 実際, 京大文学部・文学研究科が伝統的にその力を発揮し続けている分野の多くは, 読めない文書を読む, そのような文献研究なのである. 本論文で提案するものは, その「読めないものを読む」という, 人文学においても最も困難な研究をサポートするツールなのである.

このことを理解すれば, 我々の計画と類似に見える, 手書き OCR や, 他のチームの手書き文書サーチ研究が目指しているものが, 我々の目指すものと大きく異なっていることがわかる. 誤解を恐れず, 一言でいえば, これらの類似研究は, 万人向けのツールを目指している. 一方で, 我々は研究者のためのツールを目指しているのである.

そして, 実現がより困難に思える「読めないものを読む研究者用のツール」の方が「素人でも使える, 誰でも読める文書用の検索ツール」より現在の技術からすると, 実現が遙かに容易であるという, 一見矛盾しているかのようにさえ思える, この事実こそが, 本プロジェクトの根幹的着想であり, また, その現実性と実効性の源なのである.

3.6 テキスト・イメージ・サーチと OCR

SMART-GS を紹介する際, 良く受ける質問が, 文字認識, つまりを使う方が良いのではないかという質問である. たとえば, 折角, このようなツールを開発しても, 手書き文字 OCR 等の技術が将来非常に発展して, 本プロジェクトの成果が不要になるということはないだろうか？もし, そのような可能性が高いとすれば, 本プロジェクトを推進する意味は低くなる. しかし, これは文献研究においては, 実は本質的にありえないことなのである. その説明をしよう.

ある文書に対して OCR が存在する（既にある, 開発可能である）と考えるとき, その文書について少なくとも次の二つの前提が仮定されている.

1. 一意性：「この OCR のこの文書に対する精度は 99%である」という言明には, その文書の読み方は一意であるということが前提されている.「正しい読み方」が存在しているから, それからの「ずれ」を読み取り精度として考えることができるのである.
2. コード可能性：OCR とはイメージとしての文書を文字コードの並びに変換するツールである. つまり, OCR はターゲットとなる言語のキャラクタのセットが確定して, それを表現する文字のコード体系が存在していることを前提としている.

一見, 当然のように見える, この二つ条件のどちらも人文学研究では成り立たないことに注意しよう.

まず, 一意性であるが, 人文学研究では, 一つのテキストの解釈が一

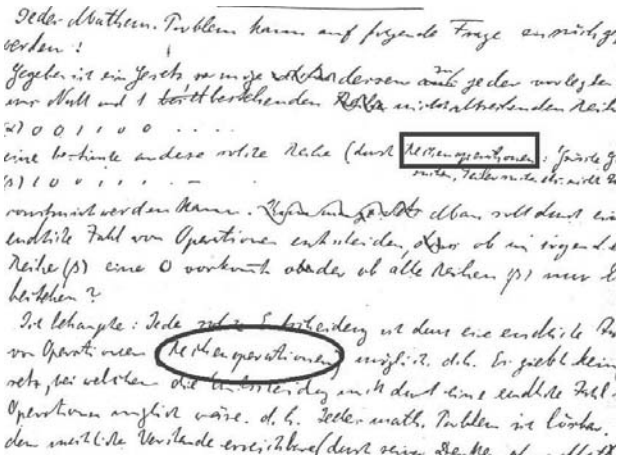

図 9: ヒルベルト数学手帳

意に決まらず, 研究者により異なる解釈が発表されることは日常的に起き得る. たとえば, 図 9 は, 林が研究しているドイツの数学者ヒルベルトの数学手帳の一部であるが, この部分の翻刻を手伝ったドイツ人学生は, これを最初, Mengenoperationen と読んだ. しかし, その後, 林が左下の楕円部分とこれが類似していることに気づき, それがコンテキストにより, Rechenoperation と読めることから, 長方形で囲まれた単語の読みが確定したという事例があった.

これは幸いにして読みが確定できた例であるが, ヒルベルト数学手帳の場合は, 読みが確定できてない例も少なくない. 幾ら手書き OCR の技術が進んでも, このような場合には適用できないであろう. もしできるとしたら, それは人間以上の数学史家としての推論能力を持つ人工知能ということになり, そのようなものが可能だとは著者には思えない.

ちなみに, SMART-GS は, 上記事例を一つの同期に開発が始められた. 今では読みにくい単語があると, それど似た例をサーチして, 似たものの読みから, 読みを推測するのは SMART-GS のユーザにとっては当たり前のことになっている. しかし, 著者が使用した感覚からすると, このツールがもっとも威力を発揮するのは, 特定の単語の出現を日記を通して見つけるというタスクである.

次にコード可能性であるが, すでに滅びた古代語, あるいは使用者が極めて少ない言語などの場合には, 文字セットの全貌が分かっていなくて当然であるから, すべての文字がコード化できていると仮定することはできない. また, 携帯電話では顔文字を一つの文字として扱えるものがあるが, これは機種に依存するわけであり, その携帯のために文字が創造されているといえる. また, 数学などでも専門分野, あるいは特定の文書のなかだけで通用する記号, つまり特殊分野のための「文字」を作ることは稀ではない. これらに OCR を使うとなれば, 文書専用の OCR が必要となってしまう.

また, 現在の OCR は読み取り精度を上げるために, 辞書や n-gram を使っているが, これは辞書や n-gram の存在を仮定しているわけだから, 辞書や n-gram が未だ存在しない言語を研究して, そららを作る研究では, OCR は使えないことになる.