

内99-19

早稲田大学審査学位論文(博士)の要旨
早稲田大学大学院理工学研究科 2956

博士論文概要

論文題目

未定義文字のコード付与方式 に関する研究

(Study on Character Code Assigning System
for Undefined Character)

申請者	
朱	青
Qing	ZHU

専攻・研究指導 電気工学専攻 情報通信網研究

1999年 11月 (工研委員会受理年月を記入)

1 研究の背景

MPEG に代表されるマルチメディア符号化技術によって、マルチメディア通信の時代が到来したと言われている現代においても、最も少ない情報量で人間の意志を明確に伝えるメディアは『テキスト』であり、そのテキストを形成しているのが『文字』である。各種携帯端末が存在していても、人間は紙上に印刷された『書類』を手放すことができない。その意味で、人類の歴史において最も重要な発明は『文字』と『紙』であると言っても言いすぎではないだろう。

人類史上初めての文字とされるのは、紀元前 3 千年頃のメソポタミアのシュメール人が絵文字を記号化して創った楔形文字だとされている。それから数千年の歴史を経て、『文字』は『一つの社会的な取り決めとして認知された言葉を表す記号』として、世界中に様々な文字が存在している。一方、『紙』が発明されたのは、紀元 100 年頃、漢の時代の中国と言われているが、『紙』の登場によって、人類の情報通信環境は一変した。また、手紙による通信が可能となつばかりか、書物によって歴史を後生に伝える手段を手にした。中国の活字と活字印刷の発明によって『本』が誕生し、1455 年のグーテンベルグによる活版印刷技術によって、大量印刷が可能となった。印刷(文書作成)技術は、1874 年のタイプライター(欧文)の発明、更には 1977 年の日本語ワードプロセッサの発明と続き、現在はコンピュータの小型化、高性能化、プリンタ技術の発達によって PC 上で高品質な文書の作成、印刷が容易に行える DTP(Desk Top Publishing) の時代に入っている。

19 世紀に入ると、1835 年のモールスによる電信機、1876 年のペルによる電話機の発明によって電気通信技術の歴史が始まり、文字による通信を瞬時に行いたいという要求から電報による通信環境が整えられた。そして、現在でも文書通信に頻繁に用いられているファクシミリの原理は 1843 年のペインによって発明されている。

文書作成端末が通信ネットワークと接続して本格的に使用されるようになったのは、インターネットの前身である ARPANET(Advanced Research Project Agency) によるネットワーク構築(1969 年)以降である。現在、高性能 CPU を搭載した PC とインターネットによる電子メールによる文字通信が日常化している。

しかし、人間の欲求を完全に満足する文字通信は未だに実現されていない。それは、文字通信が『文字コード』という、規格機関によってあらかじめ定められたコードに含まれる文字のみに対応していることにその原因がある。現在までに、人類によって作られた文字種は、約 400 程度を数えるが、そのうち電子メール通信が行えるのはごく僅かである。現在の通信環境下で、電子メール通信が行えない文字を本論文では『未定義文字』と定義する。未定義文字には、既に使用されていない古代文字の他、現在も少数民族に使用されている文字が多数含まれており、眞の意味での『グローバル』な文字通信環境を実現するためには、現在の文字コードの作成手順を含めた見直しを行う必要があると考えられる。

2 研究の目的

コンピュータで文字を扱うことは普通のことになってきたが、それとともに外字の問題、字体の問題などが生起した。送信した文書中の文字が他の記号や文字に置き換わる、におきかわるいわゆる“文字化け”も頻繁に起こっている。さらに、インターネットのようなグローバル情報通信網と航空機に代表される交通機関の発達に伴い、人間の国境、海を越えての移動が頻繁とあり、外国語による通信が日常化し、システムのマルチリンガル化は重要な課題となり、複数の言語に対応したブラウザやエディタ、ワープロも開発されている。マルチリンガル環境を実現するためには、文字フォントの互換性を実現することが最も重要な課題であり、この問題解決方法として生まれたのが『Unicode』である。『Unicode』標準は漢字圏を含めて世界的規模での共通コード体系を求めた結果定められたものであり、半角文字も全角文字も全て 2 バイトで表現することによって、漢字や英数字の区別なく、全ての文字を同じように扱うものである。そのため、英語版のオペレーティングシステムでもフォントや日本語入力プログラムを追加するだけで、日

本語が利用できるが、数々の問題点が残されている。そのうちの一つが、『漢字統合』の問題である。日本だけでなく中国や台湾、韓国でもそれぞれ独自の漢字コード体系を使用しているが、これらの漢字全てを収録すると、2 バイトでは足らなくなってしまうのは現実である。そこで、『Unicode』では、日本、中国、台湾、韓国で使用されている漢字のうち、形の似た字は同じ漢字として扱う(同一コード値を割り当てる)ことにしておいたため、自国の文字環境を守りたい漢字圏の各国から反発を受けている。その他にも、『Unicode』には既存の ASCII コードとの互換性が殆どないという問題も指摘されている。

一方、パーソナルコンピュータやインターネットの普及による情報化の進展とともに、多種多用の文字を使用したいという要求が高まっている。例えば、漢字を始めとする各国の多字種文字の扱いや、古代文字に関する標準化の問題が検討され、新たな標準案作成の動きもある。また、漢字のような同一字種の中でも、地名や人名に代表される多くの漢字の追加要求がある。これらの新規文字の追加要求は際限なく存在しているように観察されるのにもかかわらず、現在の標準化の枠組みは、必ずしもこのように多様化するニーズを満たしたものとなっていない。そのため、新たな文字コードを制定した直後にまた新たな国際標準が必要になる状況が生まれることが十分予想され、いわゆる文字コードの『青天井』問題に対する対策が見つかっていない。

文字コードはコンピュータ上で、文字や記号を扱うために、各々の文字や記号に割り振った符号であり、現在の文字コード体系は各国の規格機関で、国の一の代表者に閉じた作業で作られたものである。例えば、ASCII コードはアメリカ国内のコード体系であり、JIS 漢字コードは日本国内のローカルなコード体系に過ぎない。これらは、各自国の事情によるある順序で、例えば、“あいうえお”順とか、アルファベット順とか、各国の政府と組織による作られてたコードと記号間の対応表である。しかし、規格機関がない国や、少数民族の文字、古代文字など、コンピュータ上で今まで扱うことはできず、『未定義文字』として取り残されてしまっている。このような文字はインターネット上の情報交換に用いることができないため、国際的な地位はますます低下し、いずれは、消え去って行くことになり兼ねない。現在の文字コード体系では、大量の人力を投入しても、前述の『未定義文字』を現行の通信環境下で取扱えるようにする、すなわち『定義文字』化するには、時間がかかり、また経済性も高くないため、このまま『未定義文字』のままとどまってしまうだろう。

このような背景から、本論文では、多様な文字を用いた通信を可能とする次世代型通信環境を構築するための研究を行うことを目的としている。具体的には、文字の構造特徴について考察し、パターン認識技術によって文字を構造解析し、その構造特徴を記述する方法について研究を行う。更に、未定義文字を定義文字化する未定義文字処理システム(UCPS—Undefined Character Processing System)を提案し、UCPS を用いる文字コード付与センター(CCAC—Character Code Assigning Center)の設置を想定した文字コード付与システムの作成を目指す。

すなわち、世界中の文字をその文字の構造に基づいて記述することで文字コード付与を柔軟に行う仕組みが確立され、新たな文字にコードを付与することが可能となる。

3 本論文の構成と概要

本論文は本文の 6 章で構成されている。以下、第 1 章～第 6 章の概要と各章の関係について記す。

第 1 章 序論

本章では、本研究の背景と目的、及び各章の概要と各書の関係について述べる。

第 2 章 文字の構造解析手法の提案 1—モルフォロジーを用いる漢字の構造特徴に基づく解析法

本章では、同じ漢字文化圏に属し、歴史的に見て多くの文化的類似点を持つ日本・中国・韓国の 3 カ国における文書通信の相互接続システムが提案されている。このシステムを実現するためには漢字構造解析が必要である。従来の漢字構造解析手法は、直接基本ストロークを対象とするので、線

幅が変化し、曲線で構成されている文字への対応は困難である。本手法では、新たなモルフォロジー解析方法を導入し、モルフォロジーの形状フィルタ特性について検討したうえで、漢字の構造解析に有効な二次元複雑構成要素による方法を採用した。また漢字の構造特徴を分析し、“部首要素”と“部品要素”を選ぶ方法および解析手法について検討した。この“部首要素”と“部品要素”による解析方法は、従来の手法では解析しにくい文字にも対応しうることを示す実験結果と評価について述べる。

第3章 文字構造解析に基づく未定義文字処理システムの提案

本章では、端末上でフォントや文字コード体系が用意されておらず、その入出力環境が備わっていない文字（未定義文字）を対象とした文字処理システム（未定義文字処理システム-UCPS（Undefined Character Processing System））を提案する。未定義文字としては古代文字、少数民族の文字など歴史的に入出力環境が実装されていない文字や、特定の言語システム環境下でのそれ以外の文字も含まれる。本提案システムでは、これらの未定義文字を対象として、システムの自動化、処理時間の高速化、フォントの制約条件からの独立、文字の再構成をシステムの設計の狙いとした。本章では、モルフォロジーを用いる文字構造解析の改良法を提案している。その結果に基づいて文字の形状を記述するコードを生成し、またこれらのコードから文字を再構成方法について述べる。また、システムを試作し、未定義文字の構造解析・再構成実験を行った結果と評価について述べる。

第4章 文字の構造解析手法の提案2—輪ゴムかけに基づく文字輪郭特徴情報の記述による解析法

本章では、第2、3章で提案した手法の問題点、すなわち文字構造解析の高速化と文字記述の容易化に対する解決法を検討した。文字は、素材（图形記法）を規則（筆順）で組み合わせたものである。人が文字の筆順を勉強し、文字を読み書きできるのと同様に、コンピュータも素材（コンピュータに利用できる部品图形）と規則（コンピュータ筆順）を用いて、文字を認識する機構を実現できる。本章では、輪ゴムかけを導入し、このコンピュータ筆順の概念と実現手法を提案した。コンピュータ筆順を作成するため、輪ゴムかけに基づく、2値化された文字画像に対して、文字の特徴抽出・記述する方法について述べる。文字の輪郭情報に着目し、視覚的な凹凸情報を抽出し、文字の形状を段階的に表現する手法について述べる。本提案手法においては、概念的に文字に輪ゴムをかけ、輪ゴムと文字輪郭線との接触/非接触関係を記述する。提案法を用いて、モルフォロジー方式と比較し、各種の実験内容と結果及び評価について述べる。

第5章 未定義文字処理システムを用いる文字コード付与システムの提案

本章では、既存文字コード体系において、現在注目されている問題点に対して、UCPSを用いる文字コード付与システムを提案した。第4章で定義したコンピュータ筆順により文字を記述し、文字コード自動付与システムの作成を検討している。提案システムでは文字コード付与センター(CCAC-Character Code Assigning Center)を設置し、ユーザの新規文字登録請求に応じて、文字のコードへの登録を行い、文字コードをアップデートする。ユーザは必要に応じて、新たな文字コードブックをダウンロードして環境設定をすることで、その文字が使える環境を整えることができる。新規文字登録の判断には、未定義文字認識システム(UCPS)が用いられる。さらに、個人コードブックと共通コードブックの概念を定義し、文字IDコードの作成及びモデル実験内容と結果について述べる。

第6章 第2章～第5章までの成果を総括し、今後の課題について述べている。また、未定義文字処理システムを用いる電子図書館と文字フォントデータベースの構築への応用について述べる。