平成10年度プロジェクト研究最終発表

「文字コード自動認識機能付き ファイル・ビューワの開発」

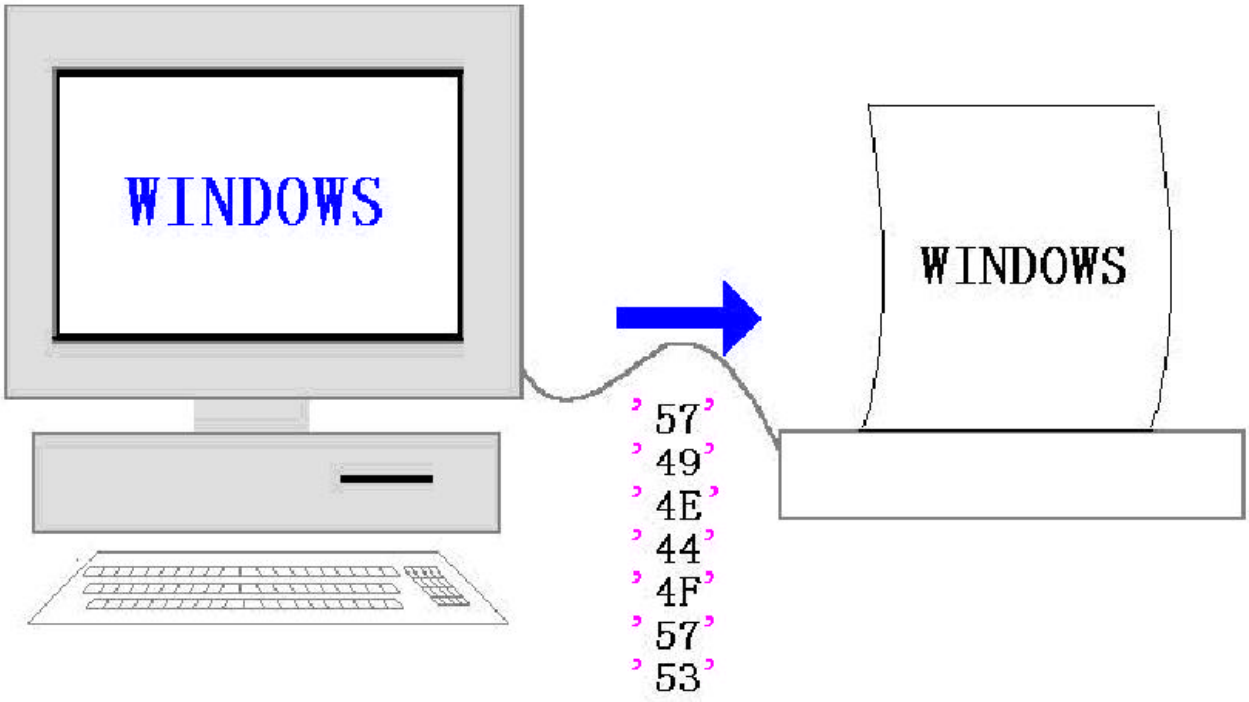
担当教官　大岩幸太郎　教官

情報教育コース
0710300　永岩里恵

【目　次】

1. はじめに

コンピュータの内部
文字を数値として処理する 符号(コード)化」
対応する数値によって文字を特定可能。



日本語の文字コード

- JIS コード
- シフトJIS コード
- 日本語 EUC コード

2. 日本語の文字コード

2.1 JIS コード

Japanese Industrial Standard の略。7ビット符号化方式。
エスケープシーケンスを用いて1バイトモードと
2バイトモードを切り替える。

図：エスケープシーケンスを用いてのモード切り替え

表：JIS コードの仕様

2バイト文字	
第1バイト範囲	21 - 7E
第2バイト範囲	21 - 7E
JIS7 半角片仮名	
バイト範囲	21 - 5E
JIS8 半角片仮名	
バイト範囲	A1 - DF
ASCII/ローマ字	
バイト範囲	21 - 7E

2.2　シフトJIS コード

Shifted　JIS　Code　の略で、8ビット符号化方式。
JIS コードとの衝突を避けるためにシフト（移動）した。

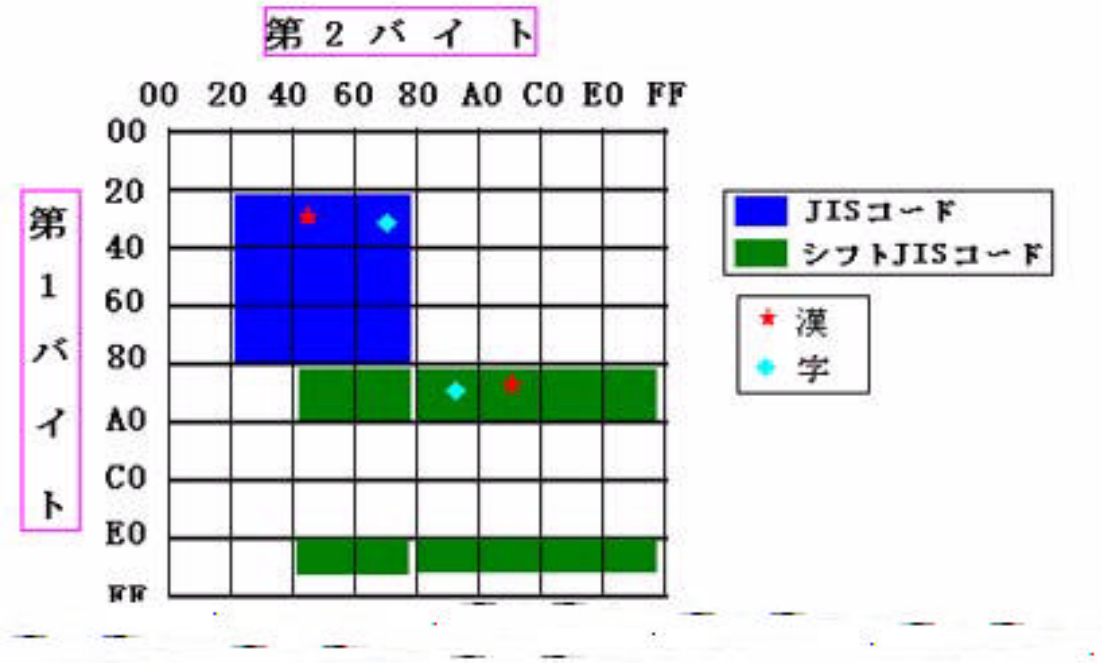


表 :シフトJIS コードの仕様

2 バイト文字	
第 1バイト範囲	81 - 9F,　E0 - EF
第 2バイト範囲	40 - 7E,　80 - FC
半角片仮名	
バイト範囲	A1 - DF
ASCII/ローマ字	
バイト範囲	21 - 7E

2.3　日本語 EUC コード

Extended　Unix　Code の略。8 ビット符号化方式。
マルチバイトコードをサポート。 日本語の符号化専用ではなく日本語や、それ以外の言語を処理するため開発された。

日本語 EUC
- EUC 圧縮フォーマット　（主流）
- 2バイト固定フォーマット

表：EUC 圧縮フォーマットの仕様

コードセット0 (ASCII/JIS ローマ字) バイト範囲	 21-7E
コードセット1 (JIS コード) 第 1 バイト範囲 第 2 バイト範囲	 A1-FE A1-FE
コードセット2 (半角片仮名) 第 1 バイト範囲 第 2 バイト範囲	 8E A1-DF
コードセット3 (補助漢字) 第 1 バイト範囲 第 2 バイト範囲 第 3 バイト範囲	 8F A1-FE A1-FE

3. ソフトの概要

・テキストファイルを開く際、日本語文字コードを検出し
何コードかを表示する。

<文字コードの検出>

ファイルを1バイトずつ読んでいく。
読み込んだバイトが ESC （1B ）であれば
⟶ JIS コードと判定

その他のコードの検出は文字のバイト範囲によって判定。

図 ：JIS コードのファイル

図 ：シフトJIS コードのファイル

図 ：日本語 EUC コードのファイル

4. 研究結果

開発環境

機種	COMPAQ　DESKPRO
OS	Windows98
言語	Borland C++ Ver5.0

開発結果

実行 ファイル名	実行ファル サイズ
fileviewer.exe	419KB

5.参考文献

(1) Ken Lunde 著

「日本語情報処理」　1998
ソフトバンク株式会社

(2) Duke 著

「Borland C++ 達人テクニック」(上・下巻)　1995
株式会社ディー・アート

(3) Clayton Walnum 著

「BORLAND C++ 4.X プログラマーズハンドブック」　1996
株式会社ビー・エヌ・エヌ

参考サイトURL

(1)日本語と文字コード

http://www.kanzaki.com/docs/jcode.html

(2)文字コードの国際規格について

http://turbine.kuee.kyoto-u.ac.jp/FAQ/kanji-code.html

(3)インターネットメールの注意点

http://www02.so-net.ne.jp/~hat/imail/cover.html