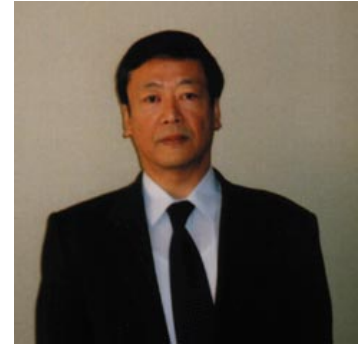


Study for Multilingual Text Processing
on the Multimedia Communication System
マルチメディア通信システムにおける
多国語処理の研究

プロジェクトリーダー 田 村 毅

東京大学

大学院人文社会系研究科 教授



Μῆνιν ἄειδε, θεά, Πηληϊάδεω Ἀχιλῆος
οὐλομένην, ἣ μυρὶ Ἀχαιοῖς ἄλγε' ἔθηκε,
πολλὰς δ' ἰφθίμους ψυχὰς Ἀΐδι προΐαψεν
ἡρώων, αὐτοὺς δὲ ἐλώρια τεῦχε κύνεσσιν
οἰωνοῖσί τε πᾶσι, Διὸς δ' ἐτελείετο βουλή,
ἐξ οὗ δὴ τὰ πρῶτα διαστήτην ἐρίσαντε
Ἀτρεΐδης τε ἄναξ ἀνδρῶν καὶ δῖος Ἀχιλλεύς

祇園精舎の鐘の聲 諸
行無常の響あり 沙羅
雙樹の花の色 盛者必
衰の理をあらわす お
ごれる人も久しからず
唯春の夜の夢のごとし

1. 研究の目的

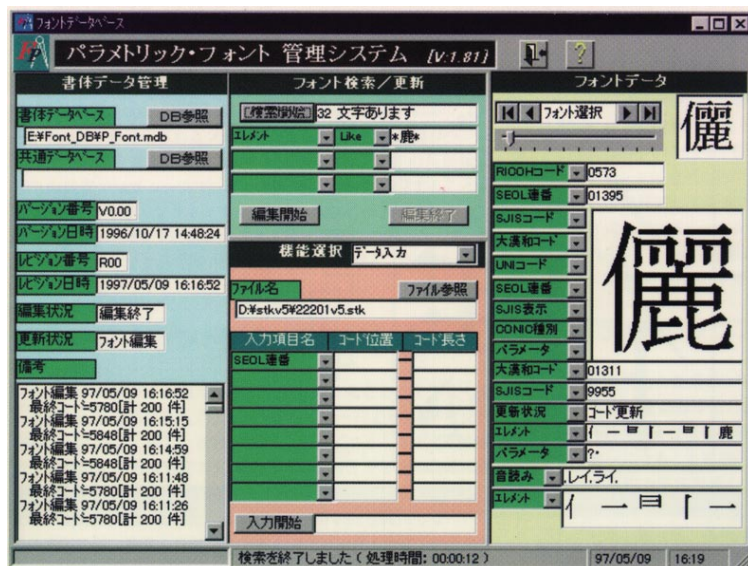
上の例のように、古代ギリシャ語（ホメーロスの叙事詩）と日本語（平家物語）のテキストを比較するような多国語処理の作業は、文学研究では日常茶飯事です。この例の場合は、或るOS上の或るアプリケーションで同時に表記して操作することが可能になっていますが、これは偶々そのOSは古代ギリシャ語フォントを実装することができ、平家物語の冒頭では、これも偶々、JISで規定された漢字以外は使われていないからです。けれどもひとたび通信となると、この程度の文字情報（古代ギリシャ語と日本語の混在する文書）ですら、そのまま誰にでも伝えられるという訳にはいかなくなります。古代ギリシャ語はひとまず措くにしても、日本語においてですら、JIS規格に含まれない漢字は、端末の機種毎に異なる「外字」として少数装備されているに過ぎません。しかも「外字」では同一性は保証されないのです。文字を画像化して通信することは、通信量の飛躍的増大につながるだけでなく、文字情報と画像情報の質的相違に目をつぶることになります。

このプロジェクトは、上に述べたような現状をふまえて、あらゆる文字に番号（コード）を与えることによって、多国語の文字情報が文字情報として自由かつ容易に通信される基盤を作ることを目的としています。

2. 研究の内容

まず日本語の漢字について、6万4千字を同定し、それぞれに番号を付けた台帳を作ることが、当面の目標となります。字数は細かい差違まで区別して「異体字」を収録していき、最終的には8～10万字に及ぶと予測しています。字形の確認と同定のために、既存のフォントでは字数も少なく、著作権上も自由に引証することはできないので、同定した漢字の字数分だけ、どうしても独自のフォントを作る必要があります。こうしてフォント・メーカーに作成委嘱しているのが「GT明朝体」と名付けたフォントです。大量の字形を短期間で作成するために、漢字を最小単位の構成要素（エレメント）に分割し、この構成要素を独自デザインで作成することにより、構成要素の組み合わせで作られたフォントも、このプロジェクトの独自デザインとなるようにしています。

こうして作成したGT明朝体フォントで字形を確認し、他のコード系のコードとも対照できるよう、漢字データベースへと編成していきます。このデータベースが整備されれば、既存システムからも参照することで、システムに依存することなく文字情報の通信が可能となりますが、それだけでなく、従来までの音訓・部首・画数だけでなく、曖昧な字形や字形の一部（構成要素とその組み合わせ）からも、文字が検索できることになります。



フォント作成管理データベース画面

3. 研究の体制

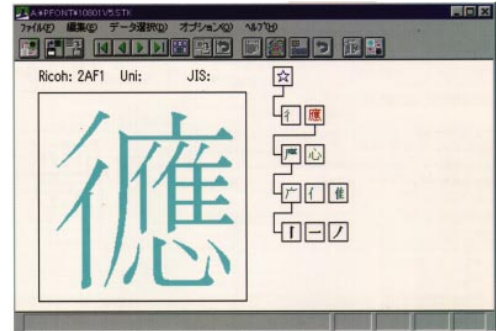
期 間：1996年10月～2001年3月

構 成：プロジェクトリーダー 1名
コアメンバー 2名

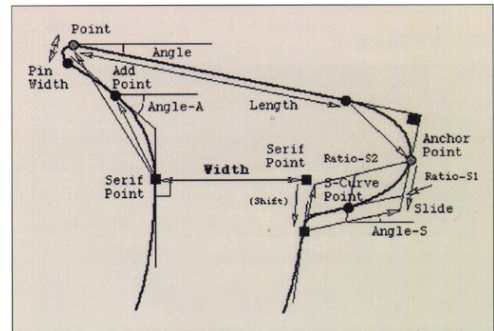
研究協力者 9名

実施場所：日本学術振興会直営研究室

：東京大学大学院人文社会系研究科 仏文研究室



漢字の構成要素への分解



エレメント・パラメータの例

この作業と並行して欧米・アジア諸国で整備されてきている各種コード系に呼応して、多国語処理の統一を検討していきますが、まずは日本語の漢字を正しく情報メディアで文字情報として伝えられるようにしなければなりません。



G T明朝体漢字フォントの例