電子化テキストの国際的共有

電子化テキストの国際的共有—付総括—

豊島正之

キーワード：文字コード、JIS漢字、10646-1、SGML、TEI、電子化テキスト（e-text）

要　旨

平成5年度国語学会秋季大会（北海道大学）での、研究発表会分科会Cセッションの概要を紹介し、そこで取上げられた下記の分野の問題と、附随する公開の問題に就て、その後の進展を加味して要約する。

1. 文字コード
閉じた文字集合を暗黙に予期している現在のコード化文字集合は、ISO 10646の様な、漢字を含む国際共通文字コード系が制定された現在でも、漢字系の様に開いた文字集合の為には、不備である。

2. テキスト形式
SGMLの難点は、階層的テキスト構造表記しか無い点と、範囲指定・属性付与を共用する設計であり、SGMLに載ったTEIの様なテキスト交換形式も、その問題を受け継いでいる。

1　セッションの概要

以下では、「電子化テキスト」（機械可読 machine readable テキスト）を簡単にe-textと言う。

従来のe-textは、特定メーカの機械、特定ソフトウェア（例えばデータベース）が前提になるなど、計算機環境（ハードウェア、基幹ソフトウェアのOS）に強く依存したものが少なくなかったがこれは、計算機環境が多言語を自由に扱うには余りに非力だった為である。計算機環境の国際化対応が一斉に進んだのは、ここ数年の事に過ぎない。

こうした環境面での国際化と、国際コンピュータネットワーク the internet の爆発的な拡大に伴い、e-textの国際的共有も現実的になって、既にe-textの公開・頒布・検索サービス（全て無償）を世界的に提供する大学・研究機関も多数現れている。

分科会C「電子化テキストの国際的共有」は、この様な世界的趨勢を踏まえて、主として日本語の電子化テキストを国際的に共有する為に、交換の為の環境整備の問題と、実践報告とで構成する事とした。