Unicode 正規化とは、Unicode の仕様の一部で、同じ文字を表す方法が複数あるときに、ひとつの方法に揃えることを指します。

Unicode には、アクセント記号等ダイアクリティカルマークのついた文字が多数あります。こうした文字の符号化の考え方として、ダイアクリティカルマークのついた形で収録することと、ベースの文字とダイアクリティカルマークとを別々に符号化して組み合わせることとがあります。

前者は例えば ISO/IEC 8859-1 に見られます。同規格には a, e, o などのアルファベットに対して、アキュートアクセントやウムラウト、サーカムフレックス等のダイアクリティカルマークがついた文字が収録されています。Unicode はこれらをそのまま収録してもいますが、一方で、後者の考え方による合成用のダイアクリティカルマークも用意しており、基底文字 (通常のアルファベット) に結合文字を後置することで、記号のついた文字を表現します。

日本語の平仮名・片仮名の濁点についても同様で、「が」という字を表すのに、単一の (合成済みの) 符号位置で表すのと、「か」に合成用濁点を後置するのと、ふたとおりの符号化表現が可能です。

同じ文字に対する符号化表現が複数あるのは不都合なので、いずれかの方法にそろえる方法が正規化として Unicode 仕様に含まれています。

Unicode の正規化には 4 種類あります。NFC, NFD, NFKC, NFKD の各形式です。名前に「C(omposition)」のつく形式は、合成済みの符号位置に極力そろえるもの、「D(ecomposition)」は分解した表現にそろえるものです。「K」のついたものは、(いわゆる) 全角英数字を通常の英数字にそろえたり、あるいは丸付き数字をただの数字に変えるなどの変更を行います。

CJK 互換漢字は、いずれの正規化形式でも、対応する CJK 統合漢字に置き換えられます。

関連項目

- CJK 互換漢字
- ダイアクリティカルマーク