L2/04-038

Status of the Unihan Database

John H. Jenkins
Apple Computer, Inc.
28 January 2004

Regarding the status of the data fields in the Unihan database, all but the following are complete: kFenn, kHKGlyph, kMeyerWempe, kRSJapanese, kRSKanWa, kRSKorean, kSemanticVariant, kSpecializedSemanticVariant, kTang, and kZVariant. (Between 4.0 and 4.0.1, the kCihaiT, kCowles, kGradeLevel, and kLau fields were completed.)

Of the complete fields, eight are normative, namely the eight IRG source fields (kIRG_GSource, kIRG_HSource, kIRG_JSource, kIRG_KPSource, kIRG_KSource, kIRG_TSource, kIRG_USource, and kIRG_VSource). The remainder, I think, can be labeled as informative, in the sense that the data in them has a good degree of accuracy and is sufficiently complete for every-day use, the main gaps which remain being in Extension B.

The exceptions are four of the reading fields, the two Japanese reading fields and kKorean and kCantonese.

In the case of kCantonese, it's been reported to us that there are systematic problems in the readings for Extension A regarding the two vowels A and AA. In point of fact, these two vowels are quite close and various authorities will not infrequently disagree on which one a particular character has, so it's not a terribly serious problem to begin with. In fact, it's a slight enough problem that I feel we can make the field informative.

As for the other three reading fields, they mainly need to be checked for consistency in the romanizations. If someone has code that I could use (preferably in Perl), for example, to convert the romanizations we use to kana and hangul, we can at least make sure of that and add these to the informative fields.

Of the incomplete fields, kMeyerWempe is somewhat over half-full, and kHKGlyph roughly one-quarter full. I'm filling these in in my spare time, and when I'm done will move on to kFenn and kTang. There are approximately five thousand entries possible for kFenn and perhaps 3500 for kTang.

Filling in the three remaining RS fields will require getting copies of the appropriate dictionaries and going through them. Fortunately, with the IRG official Japanese and Korean dictionaries, this isn't as bad as it sounds since we have indices we can use into them. An authoritative Japanese dictionary which uses the KangXi radical system and is *not* Morohashi would be good for the kRSJapanese field.

A start can be made on the three variant fields by using the dictionary indices. Since Mathews and Meyer-Wempe both use the same index for multiple characters if they mean the same thing, we can algorithmically look for such cases and use them as a core from which to build. (We'll have to cull the results for the Z-variants.) Just as a quick experiment, doing this with the kMathews field raises the number of characters with semantic variants from 150 to 2287.

Because Cora is no longer available half-time to focus on Unicode work, but needs to do it when she isn't needed for work for Apple, I'd like to have her focus on generating CDL descriptions for our ideographs, because that's the data we need most desperately.