IRG N2656

Subject: Proposal for Unique IRG Source References

Date: 2023-12-13

Author: Michel Mariani

This proposal is intended to correct the issue of duplicate IRG source references (`kIRG_GSource` property) in the **Unihan_IRGSources.txt** data file.

```
"G4K": 3
"GBK": 86
"GCH": 247
"GCY": 66
"GHC": 553


"GZ-4321207": 2
```

Note: the specific IRG source reference duplicate `GZ-4321207` has already been corrected by the UTC after validation by the IRG (177-C6), and therefore is not included in this proposal.

Current Issue

- IRG source references are normally designed to be unique, and duplicates are not allowed (see IRGN2648).
- There are currently as many as 955 duplicates made of only three characters, among: "G4K" (3), "GBK" (86), "GCH" (247), "GCY" (66), and "GHC" (553). All of them are G-sources (since they start with the prefix "G"), and all the CJKV characters they are assigned to belong to the block 「CJK Unified Ideographs Extension B」.
- Except for these duplicates, all other IRG source references are made of a short alphanumeric string followed by some kind of serial number, separated by a '-' (dash) character. It should be noted that those irregular shortened strings are still considered valid; they are specifically taken into account by the syntax regular expressions described in #UAX 38.
- This lack of uniqueness makes it impossible to provide a direct *lookup* (not *search*) feature, from source reference to code point.

Suggested Solution

- To solve this issue, it seems reasonable to modify each duplicate string by appending a dash character followed by the corresponding CJKUI's code point number, expressed in hexadecimal notation but still prefixed by the letter 'U' to prevent any collision with the current syntax format of other regular source references.
- Hopefully, these new property values follow the rules of the Annex D of the document IRGN2652.doc; they

should "blend in" smoothly in a clean and simple way, and will only require two slightly modified syntax regular expressions in UAX #38.
- Please find below the proposed modifications.

Notes:

- Obviously, the best alternative option would be to keep using the existing syntax format of the regular source references: it would require IRG experts to provide appropriate linguistic information, i.e. the actual position of the relevant characters in the source dictionaries, provided it does exist.
- China would need to agree any source reference changes.

Proposed Modifications

Updated syntax regular expressions in UAX #38: Unicode Han Database (Unihan):

```
G[013578EKS]-[0-9A-F]{4}
| G4K-(\d{5}|U[0-9A-F]{5})
| G(DZ|GH|RM|WZ|XC|XH|ZH)-\d{4}\.\d{2}
| G(BK|CH|CY|HC)-(\d{4}\.\d{2}|U[0-9A-F]{5})
| GKX-\d{4}\.\d{2,3}
| G(HZ|HZR)-\d{5}\.\d{2}
| G(CE|FC|IDC23|IDC|OCD|XHZ)-\d{3}
| G(H|HF|LGYJ|PGLG|T)-\d{4}
| G(CYY|DM|JZ|KJ|XM|ZFY|ZJW|ZYS)-\d{5}
| GFZ-[0-9A-F]{4}
| GGFZ-\d{6}
| G(LK|Z)-\d{7}
| GU-[023][0-9A-F]{4}
| GZA-[123467]\d{5}
```

Updated entries in Unihan_IRGSources.txt:

See the *kIRG_GSource.txt* data file (PDF attachment).

(End of Document)