Product: H&J

Enhancement: **New meta-word hyphenation**

Description of problem: PubLink's hyphenation program, which combines a word-breakpoint dictionary with language word-breaking logic and user's controlling rules, performs hyphenation of alphabetic words very accurately. However it was not designed to deal with longer heterogenous "words" composed of letters, digits, symbols, punctuation, and changes in type-style and baseline (eg. subscript numbers). The prominent examples of such longer strings are chemical and drug names. The program did not provide any solution to heterogenous words, other than the [ch command allowing user to insert manual breakpoints.

Solution: With Rev 16.00, PubLink includes meta-word hyphenation in all systems. A "meta-word" is a super-word, that is, a character string of unlimited length which is composed of alpha words and numbers joined together by groups of punctuation, bracketing characters and symbols. Characters may switch between roman/italic/bold, between fonts, and may change size and shift baseline. Style changes and symbols within a meta-word may happen by direct command, autosort, format call or tag.

The program determines breakpoints in a meta-word by a combination of rules. It first parses a meta-word into zones of three types:

- Alpha zone: Simple hyphenable alphabetic word.
- Symbol zone: Any group of heterogenous characters not including separator punctuation and brackets. It may include digits, letters (max three consecutive), greek letters, other symbols, super/sub-script characters and changes in font and size, plus comma, apostrophe, colon and semicolon.
- Separator zone: One punctuation character or a short group of them — period, comma, colon, semicolon, slash, parenthesis, bracket and others. Separates the alpha and symbol zones from each other. Exception: Comma/apostrophe/colon/semicolon in a symbol zones remain part of that zone.

The program then applies different break logic to each zone. An alpha word goes through the normal language hyphenation procedure. A symbol zone sets breakpoints around certain characters according to precedence rules, such as: After comma, before embedded cap, after subscript characters. In a separator zone, hyphen can go before an open paren/bracket or after a close paren/bracket (except if the separator falls at end of the meta-word); otherwise the meta-word will not break around the separator.

How to use: By default, documents use traditional hyphenation. To activate meta-word hyphenation in a document, issue the command: `[dh,metaword]`. To return to traditional hyphenation, do: `[dh,english]` (or foreign-language logic name other than `english`). In either case, a dictionary-group name may be stated before the comma, but if it is not given, the comma must remain and the default dictionary group will remain in effect.

Operational details: For documents containing no meta-words (as defined above), it should make little difference which logic you use. However the meta-word logic does do a few things differently even for simple alpha words:

- Use of the conditional-hyphen ([ch) override command: In traditional logic, if [ch is used the word *will* break there, ie. all normal breakpoints are cancelled. This action prevents [ch from being pre-inserted into words predicted to break wrong; instead it is mainly useful as an editing tool after the job has been processed. Under meta-word logic,

[ch is a priority breakpoint rather than absolute override. If it falls within the 'safe' zone in which spacebands stay within their allowed range, it will be used; otherwise normal breakpoints before or after the [ch are used.

- The ''One word won't fit'' error situation is improved greatly. Traditional logic, when given a word or meta-word longer than one line, reports the error, leaves the first character on a line alone, and retries the remaining word on the next line, continuing this until the word is short enough to fit the measure. The new logic will instead break the oversize word at best spot near end of measure, with no error message.

- In traditional logic, any punctuation in middle of letters (eg. in 'samplelastname.samplefirstname') will end word-break checking until the next space or word-terminator. That string will break only before the punctuation, never after. Meta-word logic handles this properly, recognizing the punctuation as a separator and allowing break before or after.

Required data changes: For meta-word hyphenation to work properly, your keyboard layout and pentapi data files *may* have settings that need adjustment. Visual inspection of certain character recodes, and a special test program, will tell you. Verified data will prevent possible incorrect breaks around certain fixed spaces and punctuation characters.

The data in question is the 'X/Y recode' of all characters, both keyboard- and autosort-access, which defines how H&J should treat each one. If you have multiple userdata directories in different trees, the data in each such tree should be checked.

STEP 1 — visual check: In pdata (DataTool), choose "H&J Recode Tables". Open recode table 0. (If you have multiple recode tables in records 1-n, do this Step 1 for each.) Then click "Next Pg" for the second page. Examine:

- row 44, the comma. If the Zero Shift column contains 7/0/28, change it to 7/1/28.

- the ten rows 48 through 57 containing recodes for the ten digits, where the Loc column contains 65–74. If the X & Y values are 9/0 or 9/1, change to 10/0.

STEP 2 — checker program, quick to run. It checks for any Y=4 recodes, which are obsolete and should be zeroed out for Rev 16. In a shell window, just do:

```
scankeyb ⟨Tree tree-name⟩
```

The command by itself checks data files in /Penta/.default/userdata/postscript. If you have separate userdata sets in their own trees, run again for each such tree with the optional arguments `Tree` *`tree-name`*. The program writes several progress lines to the shell window, and reports any Y=4 recodes in the form:

```
Keyboard ascii val 43 '+' (loc 11) in shove 0 has X=9, Y=4      or
Autosort mnemonic ;nb in wtgrp 0 (rec 1315) has X=7, Y=4
```

Use DataTool option H&J Recode Tables to zero out "Keyboard" hits, and option "Autosort Editor" to zero out "Autosort" hits. Contact PubLink support for help.

Internal Info

Changed in Revision: 16.00

Date: Apr 7, 2008.

Customer: Merck.

Discussion: .

Programs changed: HandJ.

Datafiles changed: (none).

Source modules changed for HandJ: hnj.h hnj-structs.h hnj.c hnj-command.c hnj-init.c hy-okwrd.c hyp-english.c hyp-german.c hyp-main.c hyp-pts.c hyp-romance.c jc-flag.c jus-autosort.c jus-get.c jus-ligs.c line-def.h txfile.h