PostgreSQL y ordenación alfabética errónea con espacios y ñ

Al ordenar una tabla de PostgreSQL por apellidos, utilizando la colación «es_ES.UTF-8», se aprecian dos cuestiones, que a mi modo de ver, no son correctas. La primera de ellas es que los espacios que separan los dos nombres no se tienen en cuenta y apellidos del tipo «Martín X» acaban intercalados con los «Martínez Y», cuando lo esperable, sería ver a todos los que tienen como primer apellido «Martín» antes de los «Martínez». El segundo de los problemas es que no se hace distinción entra la letra «ñ» y la «n», se considera a la «ñ» como una «n» acentuada y, por tanto, también se intercalan entradas que tienen una «n» en una determinada posición con filas que tienen «ñ» en la misma posición. Cuando en realidad debería considerarse la «ñ» como una letra cuyo orden de colación está entre la «n» y la «o».

Al parecer, esto no es en realidad un problema de PostgreSQL, ya que éste utiliza el orden de colación del sistema operativo en el que se ejecuta [1]. Así pues, el problema, o el supuesto problema, si es que esa es la ordenación correcta, está provocado por la definición de colación utilizada por glibc en su variante «es_ES». Puesto que el fallo en la ordenación es debido al orden de colación del sistema, el mismo error se puede reproducir utilizando el comando `sort`.

Así pues, dado un fichero `lista.txt` con el siguiente contenido:

```
a aa # 01
a ba # 02
a bb # 03
ab a # 04
ab b # 05
na a # 06
nb a # 07
ña a # 08
ñb a # 09
oa a # 10
```

Si se ejecuta el comando `LC_COLLATE="es_US.UTF-8" sort lista.txt`, se obtiene:

```
a aa # 01
a ba # 02
ab a # 04
a bb # 03
ab b # 05
na a # 06
ña a # 08
nb a # 07
ñb a # 09
oa a # 10
```

Posible solución

Una posible solución es modificar la definición de la colación para la localización «es_ES». Al hilo de lo

comentado en [2], las modificaciones que habría que hacer son las siguientes.

En primer lugar, para hacer que el espacio se tenga en cuenta cuando se están ordenando entradas, se deben añadir las siguientes líneas en el fichero `/usr/share/i18n/locales/es_ES` justo antes de la línea `END LC_COLLATE`:

```
reorder-after <U00A0>
<U0020><CAP>;<CAP>;<CAP>;<U0020>
reorder-end
```

En segundo lugar, si se observa dicho fichero se puede ver como el orden de colación se importa del fichero `iso14651_t1` (que está en el mismo directorio, en `/usr/share/i18n/locales/`) y que éste a su vez, importa el fichero `iso14651_t1_common`. Para hacer que la «ñ» cuente como una letra separada, hay que hacer las siguientes modificaciones al fichero `iso14651_t1_common`.

Por un lado, comentar las siguientes líneas:

```
<U00F1> <n>;<TIL>;<MIN>;IGNORE # 357 ñ
<U00D1> <n>;<TIL>;<CAP>;IGNORE # 673 Ñ
```

Por otro, añadir las siguientes líneas, un poco debajo de las anteriormente comentadas, justo antes de las líneas que hacen referencia a la letra «o» y a la «O», respectivamente:

```
<U00F1> "<n><y>";<BAS>;<MIN>;IGNORE # 357 ñ
<U00D1> "<n><y>";<BAS>;<MIN>;IGNORE # 673 Ñ
```

Una vez hecho lo anterior, se deben volver a generar las definiciones de las locales utilizando el comando `locale-gen`. Por último, para que la nueva ordenación tenga efecto en PostgreSQL será necesario reiniciar el demonio, p.e.:

```
systemctl stop apache
systemctl restart postgresql
systemctl start apache
```

Referencias

[1] https://wiki.postgresql.org/wiki/Todo:ICU [2] https://bugs.launchpad.net/ubuntu/+source/glibc/+bug/82302